THE EFFECT OF STAGES AND LEVELS OF AUTOMATION AND RELIABILITY ON WORKLOAD AND PERFORMANCE FOR REMOTELY PILOTED AIRCRAFT OPERATIONS

THESIS
MARCH 2015

Stephen P. Katrein, 2d Lieutenant, USAF

AFIT-ENV-MS-15-M-201

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENV-MS-15-M-201

THE EFFECT OF STAGES AND LEVELS OF AUTOMATION AND RELIABILITY ON WORKLOAD AND PERFORMANCE FOR REMOTELY PILOTED AIRCRAFT OPERATIONS

THESIS

Presented to the Faculty

Department of Systems Engineering and Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Systems Engineering

Stephen P. Katrein, BS

2d Lieutenant, USAF

March 2015

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENV-MS-15-M-201

THE EFFECT OF STAGES AND LEVELS OF AUTOMATION AND RELIABILITY ON WORKLOAD AND PERFORMANCE FOR REMOTELY PILOTED AIRCRAFT OPERATIONS

Stephen P. Katrein, BS
2d Lieutenant, USAF

Committee Membership:

Maj C. F. Rusnock, PhD
Chair

Dr. M. E. Miller
Member

Dr. B. J. Borghetti
Member

AFIT-ENV-MS-15-M-201

Abstract

This thesis investigates incorporating different stages and levels of automation with varying degrees of reliability into a remotely piloted aircraft (RPA) surveillance task in order to determine how automation implementation and reliability affect operator workload and system performance. The study uses IMPRINT discrete event simulation to evaluate three levels of reliability in twelve different baseline automation implementations within a remotely piloted vehicle task. Three stages and four levels are modeled, for a total of twelve combinations, along with a baseline task with no automation. The stages modeled are the information acquisition stage, the decision and action selection stage, and the action implementation stage, coupled with the automation recommendation level, the operator consent level, the operator veto level, and the fully automatic level. The reliability is assessed at 100%, with reduced reliabilities of 80%, 70%, and 60%. This study finds that stages of automation have greater impact on performance and the workload values than levels of automation. Automation with reduced reliability is found to have significantly reduced performance for all stages except the response stage models. However, reductions in reliability are found to have little impact on operator workload.

Acknowledgments

I would like to sincerely thank my thesis advisor, Maj Christina Rusnock for her continued support, guidance, and commitment throughout my studies. I would also like to thank my committee members Dr. Brett Borghetti and Dr. Michael Miller for their support and sharing of knowledge and expertise. I would also like to thank my wife for her patience and love shown throughout my time at school.

Stephen P. Katrein

Table of Contents

List of Figures

List of Tables

THE EFFECT OF AUTOMATION AND RELIABILITY ON REMOTELY PILOTED AIRCRAFT OPERATIONS

I. Introduction

Chapter Overview

This chapter begins by covering the background of Remotely Piloted Aircraft (RPA). It then focuses on the problem of high workload in RPA operations and the solution of building automation into the system. Next, it discusses the questions of how to incorporate automation into the design of RPAs. After the questions have been presented, this chapter focuses on the best course of action to answer the questions. Lastly, the chapter addresses the assumptions associated with this research, followed by an overview of the rest of the chapters.

Background

Remotely Piloted Aircraft (RPA) have been considered as a possible alternative to manned flight for many years. The idea of having a pilotless plane was examined for operations as early as World War I. Once World War I ended, the project to develop a pilotless aircraft was discontinued in 1925 due to a lack of motivation and need for a new weapon (Van Cleave, 2003). When World War II started, the interest in pilotless planes returned and was strengthened even further during the Vietnam War. The Firebee, a pilotless plane, was one of the principal aircraft used in Vietnam "for reconnaissance, surveillance, and some electronic intelligence gathering tasks" (Van Cleave, 2003). Unfortunately, the process of gathering the intelligence from the videos took such a long time during the Vietnam War that once the intelligence was received by the troops in that area, it was usually outdated. Even so, the Firebee remained in the air, with modifications in the early 2000s allowing it to deliver payloads to the enemy. The Firebee illustrates the versatility of RPAs in their ability to adapt to changing circumstances and continues to fly to this day (Van Cleave, 2003; Gertler, 2012).

According to the Department of Defense (DOD), the rationale behind the development of RPAs falls under three situations: the "dull, the dirty, and the dangerous" (Van Cleave, 2003). The "dull" situation applies to any duty where there is a need for continuous surveillance over a certain target for a long period of time. The "dirty" situation applies to any time where the military would need to fly into areas contaminated with chemical, nuclear, or biological weapons. The "dangerous" situation applies to any circumstance where a mission poses immediate danger to flying personnel such as a close combat air support mission (Van Cleave, 2003).

RPAs, with missions such as reconnaissance, surveillance, and payload delivery, received more attention from the United States government in 2000 due to the advantages of RPAs in the Iraq and Afghanistan wars (Gertler, 2012). The United States Congress started to provide more funding for RPA conception and development, pushing the DOD to increase the pace of RPA acquisition (Gertler, 2012). As a result of the increased acquisition pace, the Predator was a rushed program and became operationally capable only 30 months after its conception stage (Van Cleave, 2003). Other RPAs like the Global Hawk and the Reaper joined the Firebee and the Predator on the battlefield, adding to the various types of missions RPAs could complete. RPA missions are not just limited to the United States Air Force; the Navy and Marines are also investigating how they could use the unique capabilities of the RPA to better complete their missions (Gertler, 2012).

In recent years, Congress has pushed for more RPAs but pilots have been in short supply due to the increased mission load coupled with declining military end strength (number of congressionally authorized personnel) (Gertler, 2012). Currently, each RPA is operated by two individuals, the first piloting the plane and the second manning the sensor(s) (Gertler, 2012). In order to continue the growth of the RPA field, changes need to be made to counteract the pilot shortage. RPA operators are being heavily recruited to ease the amount of time each operator spends flying each day. If RPA designers were able to lower the operator's workload to a level where they could control more than one RPA at a time without becoming overworked, then those operators could fly more sorties during the same length of time. Even if the reduction was slight and the operator could only take on multiple RPAs at specific times, such as the time spent flying to and from the location of interest, the productivity of a single operator would still increase.

Problem Statement

For some years now, automation has been the leading solution to the problem of high operator workload. Many different variations of automation have been attempted, with some more successful than others. The most difficult part of incorporating automation lies not in the creation of automation, but in the implementation of it. Implementing automation towards a specific goal can have a number of potential solutions, some better than others. For example, if the operator is trying to make a phone call, implementing automation could make it quicker or easier to dial the phone number.

Ways of implementing that automation could take the form of including numbers associated with names, numbers associated with buttons, numbers associated with voice recognition, or any other number of ways to aid the operator. Incorporating automation to provide the best results is the designer goal. In some cases, automation causes the system to perform worse, in which case the automation should not be implemented. Not all automation is created with the same benefits, so the designer must choose the correct benefits to build a successful system.

When dealing with an automated system, successful system performance is directly related to the amount of automation that is incorporated and the type of tasks the automation assumes. The amount of automation may affect the operator situation awareness (SA), operator workload, the results due to automation error, or a combination of these. The intent is to try and build the correct amount of automation so that the operator workload is not too high or low. The correct amount of automation will also allow the operator to have enough SA to intervene when the automation fails, and keep the system from entering an undesirable state as a result to automation error. The automation can assume many different types of tasks; however, not all tasks should be automated. If the designer can interpret the need and decide which tasks are best for automation to take over and complete, then it can be enormously helpful to the operator. If the designer creates automation to take over the wrong tasks (as deemed by the operator), then it may add even more workload to the operator. Furthermore, if automation is set to take over the wrong task, there could be disastrous results (operator errors or mission failures), thus system designers should seek to avoid this whenever possible. By attempting to define the best ways to incorporate automation into the RPAs, operators will have a system that is easier to control.

Research Objective

In order to effectively use automation, first the designer must understand the implications of their design decisions. Without an understanding of the implications, the designer can create a bad design in a variety of different ways. Those bad designs can be avoided by understanding what implementations produce the best results. By providing results for different implementations of automation to the designer, the designer will no longer have to guess at how to incorporate productive automation into the system. This research aims to provide information that can aid in the construction of automation implementation specifically in the area of RPA operations by building a discrete event simulation (DES) to assess the impacts of implementing various types of automation. The DES took the form of a collection of models within the Improved Performance Research Integration Tool (IMPRINT) to evaluate the operator workload and performance during a surveillance RPA task. The models were based off of subject data gathered from a study completed by the 711th Human Performance Wing.

Investigative Questions

In order to answer the overarching question of how automation can be implemented to aid the operator two questions need to be addressed:

1. What stages and levels of automation reduce operator workload and increase performance in the surveillance task?

Sheridan and Verplank (1978) discuss ten levels of automation, ranging from fully automated to fully manual. Parasuraman, Sheridan, and Wickens expand on Sheridan and Verplank's ten levels by crossing them with the four stages of information processing in an automated system (Parasuraman, Sheridan, & Wickens, 2000). This research incorporates those stages and levels of automation into a DES model. The model then simulates the effect that a change in stages and levels has upon the performance of the system and the workload of the operator. Six hypotheses were created to answer this question. The six hypotheses are as follows:

1) All of the automated models will have statistically significant improved performance from the baseline.
2) Each of the stages will have statistically different performance from one another.
3) As the level of automation increases, the performance will also increase.
4) All of the automated models will have statistically significant reduced workload from the baseline.
5) Each of the stages will have statistically different operator workload from one another.
6) As the level of automation increases, the workload will decrease.

2. How does the level of reliability of the automation affect the workload and performance of the user during the surveillance task?

Reliability in the automation can have a large effect on the automation's effectiveness. If the reliability is low, incorporating the automation may lead to less effective system results than a system without automation or to a potential increase in operator workload. If the reliability is high, incorporating the automation could provide assistance to the operator by increasing performance or reducing workload. This research provides an illustration of the relationship between reliability, stages and levels of automation, and two system metrics: performance and workload. Eight hypotheses were created to answer this question. The eight hypotheses are as follows:

1. Set 1 (System Performance Hypotheses)
    1) All of the models at 60% reliability will have significantly reduced performance when compared to the baseline with no automation.
    2) All of the models at 80%, 70%, and 60% will have significantly reduced performance when compared to their respective 100% model.
    3) The performance differences between stages will be significantly affected by changes in the reliability measures.
    4) The performance differences between levels will be significantly affected by changes in the reliability measures.
2. Set 2 (Operator Workload Hypotheses)
    5) All of the models at 60% reliability and above will have significantly reduced workload when compared to the baseline with no automation.
    6) All of the models at 80%, 70%, and 60% will have significantly increased workload when compared to their respective 100% model.
    7) The workload differences between stages will be significantly affected by changes in the reliability measures.

8) The workload differences between levels will be significantly affected by changes in the reliability measures.

Methodology

A DES was built using IMPRINT to model the effects of automation on operator cognitive workload and system performance. The baseline DES represented the tasks performed by human subjects enrolled in a study performed by the Human Universal Measurement and Assessment Network (HUMAN) Lab at the Air Force Research Laboratory, Wright-Patterson Air Force Base. The DES provided a continuous workload profile for the operators performing RPA tasks in a virtual environment. Human research and prototyping of automation, while producing valuable information, is expensive, tedious, and lengthy to complete. Creating a model of the human participants not only produces cost and time savings, but also permits greater exploration of alternative design options. The model was validated against the performance and subjective workload data from the HUMAN Lab experiment. The validated baseline model was then modified to model the implementation of automation on the human subjects.

Assumptions

This research is based on a previous human-in-the-loop study and thus assumes that the human participants and the task are sufficiently representative of RPA operators and operations to effectively evaluate performance and workload impacts of automation. No additional data will be collected beyond the data gathered in the study. Furthermore, the data are gathered under the assumption that the participants attempted the task with their best effort. While the participants are non-experts within a virtual environment, the performance scores and experienced workload that is contained within the model is assumed to be representative of the workload and performance experienced by current RPA operators. Due to the prior training participants received using the software and hardware relevant to the study and due to the counterbalancing used between each participant, it is assumed that no learning effects affected the data.

Preview

This chapter began with the background of RPAs and described a problem that needs to be addressed within the RPA community and solved using automation. Chapter II contains a literature review of the relevant articles, conference submissions, and theses surrounding the topics of automation, RPAs, and reliability. Chapter III addresses the first investigative question by identifying the stages and levels of automation that have the largest impact on reducing operator workload and increasing system performance. Chapter IV addresses the second investigative question by identifying the effect of various levels of reliability on operator workload and system performance. Chapter V contains a summary of the results gathered from the research as well as potential future research to be conducted as a result of this study's findings.

II. Literature Review

Chapter Overview

With the rise of more complex systems, automation has become an integral part of system success. Automation with regards to Remotely Piloted Aircraft (RPA) is a growing field, as researchers continue to advance the technology and understand better techniques to aid the pilots during flight. This chapter begins by giving a brief overview of the best way to allocate functions to machines. Next, it discusses how RPAs and automation relate to each other, followed by a discussion about the effect of automation on the operator. This chapter then explains the advantages and disadvantages of automation, which leads into the effect of different stages and levels of automation and automation reliability. The following topic is a brief history of the Visual, Auditory, Cognitive, and Psychomotor (VACP) model used to calculate operator workload within the Improved Performance Research Integration Tool (IMPRINT), which leads into the research gap that this work fills. Lastly, this chapter closes with a short conclusion on all of the topics that were discussed.

Function Allocation

Automation is contained in almost any system. As defined by Parasuraman et al., automation "refers to the full or partial replacement of a function previously carried out by the human operator" (2000), such as a calculation performed by a computer instead of a human. Automation was not always integrated into most man-made systems but when systems began to grow in scope and complexity, automation of tasks previously completed by humans became more of a necessity. In 1951, Fitts created a list comprised of six different tasks that humans performed better than machines and five different tasks machines performed better than humans, shown in Table 1 (Fitts, 1951).

Table 1: List of tasks best suited to humans or machines – adapted from (Fitts, 1951)

Humans excel in:	Current machines excel in:
Ability to detect a small amount of visual or acoustic energy	Ability to respond quickly to control signals, and to apply great force smoothly and precisely
Ability to perceive patterns of light or sound	Ability to perform repetitive, routine tasks
Ability to improvise and use flexible procedures	Ability to store information briefly and the to erase it completely
Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time	Ability to reason deductively, including computational ability
Ability to reason inductively	Ability to handle highly complex operations, i.e. to do many different things at once
Ability to exercise judgment	

This list became a cornerstone of the automation research moving forward. Although Fitts' List was created in 1951 and has been around for 65 years, it still remains a powerful tool to use when deciding on specific functions to automate. For example, the list defies the common misconception that humans should monitor systems, as Fitts explains that machines are better than humans in performing routine tasks, such as monitoring a system (Fitts, 1951). There are exceptions to that rule, but overall Fitts suggests that machines and humans have certain tasks where one performs better than the other.

RPA Automation

Although originally applied to analyze air traffic control, Fitts' List can be applied to many systems that require automation, such as RPAs. As RPAs grew in complexity, more workload demand was placed on the operators during certain phases of flight. Historically, reports of RPA mishaps in the field of 1-2 orders of magnitude higher than manned flight illustrate the importance of recognizing the cognitive demand placed on the operators (Tvaryanas, Thompson, & Constable, 2006). Because of the high order of mishaps and the emerging progression towards heavier RPA use, researchers are directing their research towards developing an automated RPA system that supports an operator and reduces system errors to a minimum (Kaber, Stoll, & Thurow, 2007). One piece of research investigated a system that contains multiple aircraft for every person (De Visser, et al., 2008). By reversing the trend of relying on multiple people to fly a single aircraft, the military would greatly reduce manning costs, and reduce the stress on the current cadre of RPA operators, reducing their current work hours and permitting career advancement. Reducing the amount of required operators, whether that reduction is from two down to one or a team of three or more down to only two, requires a superior understanding of when, where, and how to incorporate automation into RPA operations.

Effect of Automation on Operator

A broad range of actions have been covered by automation in recent years, consisting of everything from dialing a number on a cell phone to an autopilot flying an airplane. While automation does relieve the human from completing whatever action needs attention, automation does not completely remove the action from the workload of the human. When automation is present, a human is usually overseeing the action performed as verification that the action is being completed. Because of this change from a worker to a monitor, the human does not fully shed the task. This causes the task to change from one form of workload to another, often resulting in a decreased amount of workload. This effect shows that automation can be useful when designers find a way to reduce workload, but researchers have yet to quantify the difference in the workload change. Consequently, understanding the new amount of tasks an operator could handle is still unknown.

Before any automation can be incorporated into the system, the system designers need to be able to identify when the automation should come into effect. If the designer incorporates too much automation, then the operator may experience underload, in which they might lose situation awareness (SA), negatively impacting performance. If the designer incorporates too little automation, then the operator workload can become excessive, again negatively impacting performance (De Visser, et al., 2008). Automation fixed problems that arose because it could control some of the more mundane tasks, but also opened the doors to a host of new problems, including issues with situation awareness, trust, complacency, decision-bias, and fluctuations in workload (De Visser, et al., 2008). To combat any tendencies towards these negative issues, the goal of a designer is to pinpoint the state where the operator is working enough to still have SA but is not overexerted to the point that performance suffers (Rusnock & Geiger, 2014). In order to pinpoint where the operator needs help, the cognitive workload of the operator needs to be captured.

Automation Advantages and Disadvantages

Automation provides some unique advantages and disadvantages. One advantage is a general reduction in human error. By moving human interaction with the system into a monitoring position, the human participation in the task is reduced (Swanson, et al., 2012). With the human slightly removed from the task, the accompanying human error is normally lessened. Also, when the automation is incorporated correctly, the overall task load of the operator will be reduced. By reducing the human's task load, the human operator is able to focus on other tasks that may improve overall system performance.

One of the disadvantages of automation arises when the human is missing vital pieces of information about the process or situation. If automation takes over every process, then the human cannot participate when the automation fails because the human lacks appropriate SA. Not only is SA lost, but reduced interaction with the system can lead to a loss of skill with regards to effectively operating the system. Automation can also potentially cause an increase in workload because of the added communication between the system and the operator. Examples of automation communication include: informing the operator of task completion, asking the operator for permission to complete an action, or asking the operator to choose between alternatives.

Trust in automation is another disadvantage that can become a problem. If the operator places excess amounts of trust in the automation, then some incorrect actions may be executed by the automation without any knowledge from the operator that the results were incorrect. If the operator places too little trust in the automation, then more time will be spent by the operator verifying or re-doing work previously completed by the automation (Cring & Lenfestey, 2009).

As mentioned above, a reduction in human error is expected when automation is implemented. Clumsy implementation of automation may, however, lead to an increase in human error (Woods, Johannesen, Cook, & Sarter, 1994). New burdens may be unintentionally placed on the operator, creating more problems and more opportunities for error, along with the expected benefits provided by the automation (Woods, Johannesen, Cook, & Sarter, 1994). For example, if automation is only built to accommodate routine scenarios, then latent problems may arise when a scenario appears that was not covered. These latent problems could then emerge when the human works through the scenario (Woods, Johannesen, Cook, & Sarter, 1994). That scenario may never occur, but the possibility of it happening leads to an added possibility of human error due to the clumsy implementation of automation.

Stages and Levels of Automation

To understand the different ways to apply automation to a system, researchers look to the human information processing model (Broadbent, 1958). The act of human information processing occurs in four stages, shown in Figure 1 (Parasuraman, Sheridan, & Wickens, 2000).

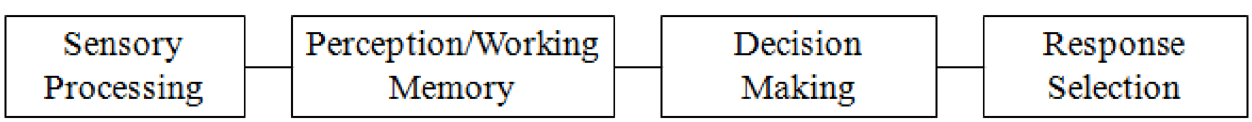


Figure 1: Human Information Processing Model – adapted from (Parasuraman, Sheridan, & Wickens, 2000)

In the first stage, Sensory Processing, the five senses gather information from the outside world and send the information to the brain. Each one of the senses receives different types of relevant information. In the second stage, Perception/Working Memory, the brain combines the information acquired by the different senses in the Sensory Processing stage with information in long-term memory to form a coherent picture of the environment. Because of the large amount of information gathered from the senses, some of the information deemed less important is not consciously perceived, or is filtered out. The Decision Making stage forms the third stage and consists of deciding on a course of action within that environment. The Decision Making stage is based on the information in the Perception stage, thus decisions may be made on incomplete information. The final stage is the Response Selection stage, which consists of completing the action decided upon in the Decision Making stage (Kaber, Stoll, & Thurow, 2007; Parasuraman, Sheridan, & Wickens, 2000)

The four stages of processing describe human decision-making, but they correlate closely with system processing as well. A system can complete the same tasks of gathering information, compiling relevant information, deciding on a course of action, and implementing that action. Based upon those similar stages, machine tasks can also be grouped into a particular stage of machine processing, leading to the four stages of automation (Parasuraman, Sheridan, & Wickens, 2000). The relationship between the two processing models is shown in Figure 2.

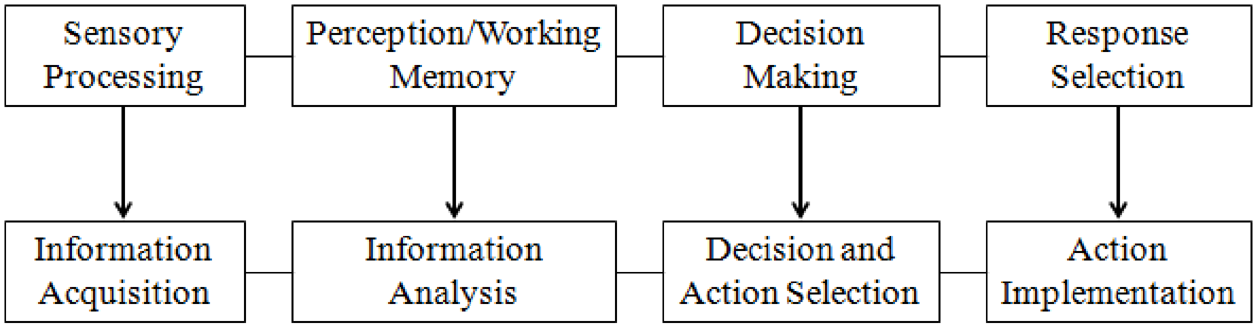


Figure 2: Comparable stages of processing models – adapted from (Parasuraman, Sheridan, & Wickens, 2000)

In addition to the four types of automation, automation allocation can also be explained by the ten Levels of Automation (LOAs), proposed by Sheridan and Verplank (1978), describe the distribution of tasks which can be allocated to either the human or the automation. The first level is considered to contain no automation because all tasks are allocated to the operator. The tenth level is considered to be fully automated, without human interaction because all tasks are allocated to the automation. The other levels contain varying amounts of automation between these two extremes. Table 2 describes the ten levels of automation.

Table 2: Levels of Automation – adapted from (Sheridan & Verplank, 1978)

	Determines Alternatives	Suggests Alternative	Selects Alternative	Executes Alternative	Informs of Action
Level 1	Human	Human	Human	Human	N/A
Level 2	Computer	Human	Human	Human	N/A
Level 3	Computer	Computer	Human	Human	N/A
Level 4	Computer	Computer	Computer, Human may or may not approve	Human	N/A
Level 5	Computer	Computer	Computer	Computer, if Human approves	N/A
Level 6	Computer	Computer	Computer	Computer, unless Human vetoes	N/A
Level 7	Computer	Computer	Computer	Computer	Always
Level 8	Computer	Computer	Computer	Computer	If Human requests
Level 9	Computer	Computer	Computer	Computer	If Computer decides to inform human
Level 10	Computer	Computer	Computer	Computer	N/A

Allowing the system designer to choose between different levels of automation within a system illustrates that automation is not just a choice between on or off, but instead exists along a continuum of varying degrees of automation. Recognizing this continuum is important because different LOAs are expected to have different effects on performance and situation awareness. For example, an LOA near the middle can improve performance and situation awareness, even as system complexity increases (Ruff, Calhoun, Draper, Fontejon, & Guilfoos, 2004). Understanding that automation resides along a continuum allows system designers to manipulate the level and stage of automation to best fit the given scenario (Cummings, Bruni, Mercier, & Mitchell, 2007; Parasuraman, Sheridan, & Wickens, 2000; Endsley, 1999).

Reliability

Reliability causes many problems for system designers. Low reliability can potentially offset helpful automation to the point that the operator's job becomes more difficult rather than less. When the less reliable automation is working directly against the goal of improving the system by reducing the performance of the system or increasing the workload of the operator, the system designer will need to make a choice to improve the reliability or remove the automation altogether.

Reliability is also partly a function of system complexity. As systems become more complex, the automation becomes more complex as well, leaving greater opportunities for unforeseen problems that could lead to a system failure. This results in the "irony of automation" where, as the complexity of a system rises, human involvement becomes more critical due to unforeseen problems (Bainbridge, 1983).

One recent reliability study in the RPA field focuses on the reliance and compliance of human dependence (Wickens & Dixon, 2006). Reliance is the state of human dependence when the automation is quiet. Compliance is the state of human dependence when the automation is alerting the human that something has potentially gone wrong. Human reliance stays high when the automation has fewer misses, meaning that the human has more trust that the system is fine when the automation is quiet. Conversely, human compliance stays high when the automation produces fewer false alarms, meaning that the human has more trust in the automation to correctly identify when something has gone wrong. When both metrics are high, the human experiences less cognitive workload because the human believes that the automation is handling the task well. Both of these metrics are based on human perception, so there is potential for a disconnect between actual automation performance and perceived automation performance. The study performed by Dixon and Wickens (2006) illustrates the reliance and compliance of the human and how those two metrics may affect the reaction time of the human to any automation signals. Dixon and Wickens found that when the automation produced more misses, the operator was quicker to notice them and fix them, but had trouble completing the concurrent tasks in a timely manner (less reliance). When the automation produced more false alarms, the operator had a slower and less accurate response (less compliance) to the alarm but showed little change in the ability to complete the concurrent tasks.

Reliance and compliance are important attributes for alarm-style automation systems; however, these attributes may be less relevant for other types of automation implementation. For example, with RPA operations, the automation may help track a target. This example does not fit in neatly with reliance and compliance which are geared towards alerts and alarms, thus reliance and compliance may be less helpful in determining the reliability of the automation. Another way to look at reliability is the percentage of time that the automation does not fail, represented as a number from 0-100% (Parasuraman, Molloy, & Singh, 1993). A failure can represent any type of action taken by the automation that the operator did not expect or any type of halt in the automation sequence, where it cannot manage to complete assigned activities. Previous automation studies have attempted to identify the point at which automation failure makes the system performance decrease and operator workload increase above the baseline of not having any automation at all. One study has placed this number at approximately 70-75% reliability (Wickens & Dixon, 2006). Thus, if the automation fails more than 25-30% of the time, then the operator would have performed better without the automation. However, the task being completed also has an impact on the effectiveness of the automation as the reliability is reduced. John and Manes found that even automation reliabilities below 70% still may be helpful (John & Manes, 2002). In their research, the goal of the operator was to locate a target while the automation would provide suggestions on places to look. As the reliability was reduced below 70%, the automation was still helpful in aiding the operator. Thus, the reliability threshold for which it begins to harm the workload and performance of the operator may depend on the task being completed. Perhaps metrics including task completion times for the human and the automation, recovery time necessary in the event of a reliability failure and operator workload could be useful in further understanding this tradeoff. System designers need to know at what threshold the automation reliability should stay above in order to help, rather than hinder, task performance.

VACP Modeling Tool within IMPRINT

In 1984, Wickens built upon the bottleneck and single resource workload theories to develop the multiple resource workload theory (Wickens, 1984). As Wickens explained, the argument for the multiple resource workload theory was that information processing required multiple resources within the brain (Wickens, 1984; Keller, 2002). These resources included the visual, auditory, spatial, and verbal among others. For example, scanning a crowd for a sibling is a task that uses visual resources. Auditory resources may be used when listening to music, attempting to understand the lyrics. We can accomplish any number of tasks at once as long as the combined information from those tasks do not overload one of the resources. Combining these two actions, listening to music and scanning a crowd for a sibling, is possible because they do not stem from the same resources within the brain. However, listening to two conversations at once becomes very difficult because the auditory channel is becoming overloaded with similar information. Building upon the basic idea of the multiple resource model, the VACP modeling tool identified four resource components: visual, auditory, cognitive, and psychomotor. These four components are each characterized by a scale of demand levels, with values assigned by a pool of subject matter experts (McCracken & Aldrich, 1984). The psychomotor channel was then broken up into fine motor, gross motor, and tactile and the speech channel was added for a total of seven channels that are being used in the DES software tool IMPRINT. This updated model is the device that captured the workload of the operators during this study and is the basis for all calculations regarding workload in this paper.

Workload and performance have been studied together before in an effort to identify what happens to the performance as workload changes (Yerkes & Dodson, 1908; Donmez, Nehme, & Cummings, 2010; Clare, Hart, & Cummings, 2010). These studies have found that when workload changes, performance is affected. The change is not linear or monotonic, and performance will peak at a certain amount of workload before it begins to decline. The amount of workload that results in peak performance seems to change as the task changes, so no specific guidelines have been able to predict performance for other tasks or other combinations of tasks.

Research Gap

Stages and levels of automation have been applied since 2000, when Parasuraman et al. explained the way that stages and levels could interact (Parasuraman, Sheridan, & Wickens, 2000). Since then, stage and levels have been incorporated into research about manufacturing systems (Johansson, et al., 2009; Sheridan, 2011) or may have focused on SA (Furukawa, Inagaki, & Niwa, 2000). In 2005, Wright and Kaber conducted an experiment that consisted of three stages of automation coupled with two levels of automation, similar to the experiment in this paper. Measures of dependent variables centered on team effectiveness and team coordination, with the results indicating that both stages and levels had different effects on teamwork (Wright & Kaber, 2005). In another experiment in 2003, the combination of another two independent variables, the level of automation and the automation reliability, was changed to measure the response of the operator (Meyer, Feinshreiber, & Parmet, 2003).

A similar experiment was conducted in 2007 (Rovira, McGarry, & Parasuraman, 2007). In their experiment, the human operator goal was to correctly select a friendly and enemy target to engage in combat. The experiment modeled two different stages of automation, three different levels for a single stage, and two different levels of reliability. While the results are not directly translatable, they do suggest that with 60% reliability, both of the stages of automation show significantly reduced performance for all levels measured.

This research aims to gather each of these research concepts together to develop a cohesive study that demonstrates the effect of changing stages and levels of automation and reliability upon operator workload and system performance within RPA operations.

While the studies mentioned above closely relate to this thesis, this thesis has a wider range of values for the stages, levels, and reliability of automation. This is possible due to the nature of DES, which allows for multiple alternative scenarios to be created once the baseline model has been built, consuming fewer resources than a human subject experiment. Much of the previous research built upon one or more of these same concepts, but few studies that combine RPA operations with different automation reliabilities, stages and levels of automation, the system performance, and the operator workload have been found.

Summary

Understanding the previous literature is a necessary step in fully understanding the problem. This chapter focused on the development of automation, the concept of workload, and the relationship between the two. The other topics discussed included topics related to the investigative questions and topics related to the tools used to create the models. Understanding the different types and levels of automation will allow for the first investigative question to be answered. The second investigative question focuses on reliability, discussed briefly in the automation section. Finally, the research around this topic was explained, demonstrating a gap that needed to be filled. The methodology will be addressed in the next chapter.

III. Modeling the Effects of Stages and Levels of Automation on Operator Workload and System Performance in RPA Operations

Abstract

This paper simulates different stages and levels of automation within a remotely piloted aircraft (RPA) surveillance task and investigates how these simulated automation implementations affect operator workload and system performance. The study uses discrete-event simulation (DES) to model the surveillance task in IMPRINT. Performance was measured based on a point system and workload was measured using the Visual, Auditory, Cognitive, and Psychomotor (VACP) model. Three stages and four levels were modeled, for a total of twelve combinations, along with a baseline task with no automation. The performance and the workload values were unaffected by the different levels of automation but were affected by the stage of automation. Automation of the decision and action selection stage produced the largest increase in performance and automation of the action implementation stage produced the largest reduction in workload.

Introduction

Remotely Piloted Aircraft Use

In the past decade, use of remotely piloted vehicles has grown significantly. As the flight hours and total number of sorties continued to grow, new challenges began to arise. In the military, only current pilots were qualified to fly the RPAs but few wanted to leave the freedom of flight to sit confined on the ground while flying a remotely-piloted aircraft. Nevertheless, the role of the RPA continued to grow through the Global

War on Terror (GWOT), Operations Enduring Freedom (OEF), and Iraqi Freedom (OIF) (Callam, 2014). Much of the focus on current and future missions is aimed at the removal of ISIS leaders, a mission well-suited to RPAs (Jones, 2014). Actions such as these illustrate the effectiveness, importance, and responsibilities that RPAs have begun to assume.

RPA use will continue to rise, but the size of the current military workforce is declining (Gertler, 2012). To keep up with increased demand, RPAs will need to act as force multipliers, multiplying the benefits without increasing demands on manpower. If additional automation can be effectively incorporated into RPA control systems, reduced workload may allow for a pilot to control multiple RPAs at the same time. Increasing the quantity of RPAs while simultaneously reducing the quantity of pilots needed to fly them can enable increased mission rates while reducing manpower costs (Taylor, 2006).

Motivation

System designers need to understand that automation consists of many possible implementations. A solution that works well in one scenario may not work well in others. The most influential automation implementation depends on the goals of the system and the system processes. When designers incorporate automation into a system, they need to consider the implications of automation implementation. This research investigates different automation options and assesses how those options impact the performance of the system and the workload of the operator within an RPA task.

Background

Automation

Automation is contained in almost any system. As defined by Parasuraman et al. (2000), automation "refers to the full or partial replacement of a function previously carried out by the human operator." Automation is typically intended to reduce task load or increase operator efficiency. Ideally, the automation allows for a balance to occur between the capabilities of the system, what the system can achieve, and the increasing demand on the human resources (Taylor, 2006).

As automation is increasingly applied to divergent or non-algorithmic tasks within systems that are employed in unpredictable environments, the human operator's tasks are not completely replaced by the automation. Instead, the operator is asked to provide supervisory control of the system and adjust the automation or assume manual control during automation failures or during operational scenarios for which the automation is not designed. As a result, the automation does not replace the operator but changes the nature of the operator's tasks, as well as the exchange of information between the system and the operator. In alternative designs, the automation and operator participate as a team, with the automation performing more mundane tasks, freeing the operator to perform tasks which require inductive reasoning or other tasks at which the human excels (Fitts, 1951).

Current designers need to incorporate automation into RPA systems in order to allow for the RPA to function without overloading the operator. For example, automation in UAVs might focus on flying the aircraft, permitting the operator to perform critical mission tasks, such as monitoring the sensor feed and deploying armaments.

Stages and Levels of Automation

As automation replaces tasks performed by the human operator, replacement may include tasks related to any of the four stages of human information processing: Sensory Processing, Perception/Working Memory, Decision Making, and Response Selection. Sensory Processing gathers information from the outside world and provides it for higher level processing. Perception/Working Memory synthesizes this information with remembered information to form an interpretation of the environment. Decision Making relies upon the interpretation of the environment to decide upon a course of action. Response Selection completes the action decided upon in the Decision Making stage. When automated, the replacement technologies are referred to as Information Acquisition, Information Analysis, Decision and Action Selection and Action Implementation, respectively, shown in Figure 3 (Parasuraman, Sheridan, & Wickens, 2000).

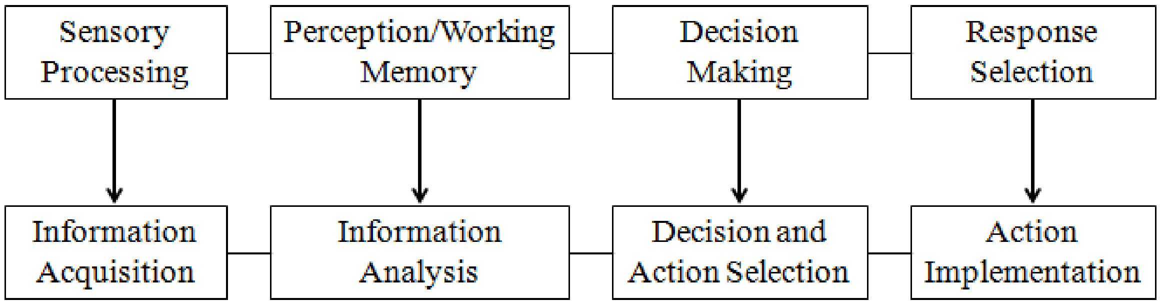


Figure 3: Stages of machine processing built from the human information processing model – adapted from (Parasuraman, Sheridan, & Wickens, 2000)

The replacement technology can automate each of the four stages of information processing along any one of ten levels of automation, as proposed by Sheridan and Verplank (1978). These ten levels of automation (LOAs) are provided in Table 3.

Table 3: Levels of Automation (LOA) – adapted from (Sheridan & Verplank, 1978)

	Determines Alternatives	Suggests Alternative	Selects Alternative	Executes Alternative	Informs of Action
Level 1	Human	Human	Human	Human	N/A
Level 2	Computer	Human	Human	Human	N/A
Level 3	Computer	Computer	Human	Human	N/A
Level 4	Computer	Computer	Computer, Human may or may not approve	Human	N/A
Level 5	Computer	Computer	Computer	Computer, if Human approves	N/A
Level 6	Computer	Computer	Computer	Computer, unless Human vetoes	N/A
Level 7	Computer	Computer	Computer	Computer	Always
Level 8	Computer	Computer	Computer	Computer	If Human requests
Level 9	Computer	Computer	Computer	Computer	If Computer decides to inform human
Level 10	Computer	Computer	Computer	Computer	N/A

The differences between these levels arise in how much responsibility the automation assumes when completing the task. These levels give system designers flexibility when incorporating automation because the levels provide a range from fully manual to fully automatic. These levels are then coupled with the machine information processing model by choosing a stage of automation and a level of automation to build a desired action. For example, Level 3 coupled with the decision and action selection stage may form an automated action that provides alternatives to a decision the operator must make. Note that in an automated system, each information processing stage can have a unique level of automation. By combining these 10 levels of automation with the four levels of processing to be automated, 40 automation combinations are available for each human task to be automated (Parasuraman, Sheridan, & Wickens, 2000; Endsley, 1999).

While these stages and levels have been used since 2000 to illustrate different automation implementations, a limited amount of research has been conducted to evaluate the effectiveness of each of these conditions on automation utility or efficiency. However, this limited research has included applications in manufacturing systems, power plant systems, or research about situation awareness (SA) (Johansson, et al., 2009; Sheridan, 2011; Furukawa, Inagaki, & Niwa, 2000). A similar experiment to the one presented in this paper was conducted in 2007, which broke the stages and levels up into two different stages and three levels (Rovira, McGarry, & Parasuraman, 2007). Ultimately, stages and levels provide a uniform way to research and study different types of automation.

While the studies mentioned above have explored stages and levels of automation, this paper explores a wider range of values for the stages and levels of automation. This is possible due to the use of discrete event simulation, which allows for multiple alternative scenarios to be easily evaluated, consuming fewer resources than a human subject experiment.

Purpose

This paper aims to illustrate the effect of different stages and levels of automation upon the system performance and operator workload and highlight any automation implementations that yield better results than others. This research will aid system designers when making decisions regarding automation implementation.

This paper addresses six hypotheses. Three hypotheses focus on operator workload and three focus on system performance. Each set of three assesses the same independent variables: the first addresses the difference between the system with no automation and the system with automation; the second addresses the difference between each of the stages of automation; and the third addresses the difference between each of the levels of automation. The six hypotheses are as follows:

1) All of the automated models will have statistically significant improved performance from the baseline.
2) Each of the stages will have statistically different performance from one another.
3) As the level of automation increases, the performance will also increase.
4) All of the automated models will have statistically significant reduced workload from the baseline.
5) Each of the stages will have statistically different operator workload from one another.
6) As the level of automation increased, the workload will decrease.

Methodology

IMPRINT and DES

A discrete event simulation (DES) model was constructed to represent an existing human subjects experiment. This model was developed in the Improved Performance Research Integration Tool (IMPRINT), a DES environment specifically tailored to model human performance. IMPRINT enables the quantitative modeling of operator workload through the incorporation of the Visual, Auditory, Cognitive, and Psychomotor (VACP) scale. The scale relies on multiple resource workload theory to quantitatively assign cognitive demand to different resource channels. The demand on each resource channel is quantified on a scale from 0 to 7, with verbal descriptions assisting in assigning of quantitative values. Overall workload can be calculated using simultaneous demand experienced by all task for all channels. Once the baseline model was built and validated, alternative models were created. These alternative models incorporated a combination of different stages and levels of automation.

Data Collection for DES Model: Human Experiment

The IMPRINT models used in this study were created using data gathered from a human subject experiment conducted by the 711th Human Performance Wing Human Universal Measurement and Assessment Network (HUMAN) Lab at Wright Patterson AFB, OH. The baseline IMPRINT model represents the subject completion of an RPA surveillance task, described below. The interfaces used to complete the task were a standard QWERTY keyboard, a right-handed mouse, a headset, and three computer monitor displays. The experiment gathered key press data, subjective workload, and performance scores. The behavior data gathered from the experiment were used to construct probability distributions which are incorporated into the DES model tasks. These probability distributions are sampled by the model to capture variability for the task times. The incorporation of the data permitted a faithful representation of distributions of task times for the human subjects in the model. Further details regarding incorporated behavior data and model validation are described in the Data Gathered section and the Generating IMPRINT Workload and Performance Values section.

Design of Human Subject Experiment

The goal of the surveillance task is to locate a high value target (HVT) walking around within a market as shown in Figure 4. In the figure, the right side shows a fully zoomed out view of the market while the left side shows a median zoom level of the market place. The HVT is carrying a rifle which differentiates it from other human figures in the environment which serve as distractors. Some distractors carry a shovel or a pistol, while others are empty handed. The operator can click anywhere on the screen to center the sensor on that position. The mouse wheel allows the operator to zoom in or out, providing the operator the ability to identify the HVT or move around the market quickly. When found, the operator presses the F key on the keyboard to begin following the HVT.

Figure 4: Screenshot of market during Surveillance Task

While the operator is completing the primary surveillance task, there is a secondary communication task that consists of answering a mathematics question. The mathematics question simulates operator communications with other pilots or air traffic controllers. The mathematics question is relayed through the headset, and takes the form of a single-step addition, subtraction, multiplication, or division problem. As an example, the operator may be asked to find how far a plane might travel given its speed of travel and a certain time period. The operator answers the problem by pressing down the space bar, and saying the answer aloud into the microphone. Both the surveillance task (primary task) and the communication task (secondary task) can be completed simultaneously.

The primary and secondary tasks in the surveillance trial are completed four times over a period of 265 seconds. Each HVT is present for 60 seconds before walking under a tent, with a new HVT appearing after the prior one has passed from view. The first mathematics question is asked 40 seconds from the beginning of the trial and subsequent questions are asked every minute thereafter. The operator has 30 seconds to answer the question, with a steady decrease in performance score as the time to answer approaches 30 seconds. The operator is unaware of the schedule of each trial and is told to continue searching for and tracking HVTs during the length of the trial. Upon completion of each trial, the operator has 180 seconds to complete the NASA Task Load Index (NASA TLX), a subjective workload questionnaire for each trial (Hart & Staveland, 1988).

The surveillance task consists of four different scenarios, intended to vary the difficulty of the primary task. The four scenarios implement two independent variables each with two levels, as shown below in Table 4. The first variable is the quantity of distractors in the market, either a high (48 distractors) or a low (12 distractors) distractor level. The second variable is the quality of the camera feed, either a high quality or a low quality camera feed. The high quality camera feed shows a clear view of the market.

The low quality camera feed shows a view of the market with visual static noise imposed over it. These two variables combine to create a total of four different scenarios. Each participant completes each scenario 4 times, in a randomized order, for a total of 16 trials.

Table 4: Experimental Design Matrix

	Low Distractors	High Distractors
High Camera Quality	Scenario 1	Scenario 2
Low Camera Quality	Scenario 3	Scenario 4

Data Gathered

Three different types of data--key press data, subjective performance data, and subjective workload data--were gathered from the study. The key press data consists of each time the F-key was pressed and each time the space bar was pressed. There was a timestamp associated with each of the key presses. The F-key was pressed by the subject every time a HVT was believed to be found. The space bar was pressed by the subject every time the subject answered one of the mathematics questions. Together with the performance data, these two pieces of data give insight into when the subject completed each task.

The performance data consists of data gathered during each second of the trial, with three points possible per second. The subject could receive a total of 800 points for the primary task and 200 points for the secondary task for a combined total of 1000 points. If the target was on the screen after the F-key was pressed, points were added to the overall score. The amount of points added to the score depended on the zoom level.

If the target was off of the screen, no points were given. Because the target was continuously moving, the operator would need to re-center the screen often to keep the target on screen. For the mathematics question, the operator would lose 5 points if the answer provided was wrong, would gain up to 50 points (depending on the length of time spent to answer the question), and would gain 0 points if no answer was provided.

The subjective workload data consists of a NASA-TLX survey at the end of each trial. The NASA-TLX provides scales for six different dimensions of subjective workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. Five of these are rated on a scale from low to high, and performance is rated on a scale from good to poor. The subjects were instructed to rate their perception on each scale during each trial. The subjective workload data is used to validate the VACP workload scores.

Experimental Design for the DES Automation Experiment

The information provided from the human experiment was used to create the baseline DES model for the surveillance task and any subsequent alternative model. In order to determine effect of implementing automation within the surveillance task, certain combinations of stages and levels of automation were chosen to be modeled in IMPRINT. Out of the forty possible combinations available to be tested (4 stages x 10 levels of automation), twelve combinations were chosen, and are described in the Automation Models section below.

The two independent variables are the stages of automation and the LOAs. The 12 selected values of these factors were deliberately chosen to capture the full range of values to ensuring substantial differences in the implementation of the automation while minimizing the number of treatment combinations. The levels of automation that were selected are levels three, five, seven, and ten, Note that level one represents the baseline scenario. The types of automation chosen are information acquisition (information acquisition stage or Stage A), decision and action selection (decision stage or Stage C), and action implementation (action stage or Stage D). The analysis stage (Stage B) was omitted from the study because at the current level of detail, this stage is combined with the decision stage and cannot be effectively separated.

The dependent variables are the performance and workload of the operator during the task. The performance is measured out of 1000 points, following the standard set in the “human-in-the-loop” experiment, with the performance averaging out to 340 points for the primary performance and 179 points for the communication performance, for a combined average of 519 points in the baseline model. The workload of the operator is determined using the VACP scores gathered from each model, producing a time-weighted average of 14.78 in the baseline model. The communication score is not included in the analysis because the secondary task is unaffected by the automation implementations.

Out of the four scenarios of the experiment, the scenario with a high amount of distractors and low camera quality was selected, thus it represented a case that is likely to benefit from automation. Scenario 4 was modeled in IMPRINT by conducting a detailed task analysis to determine the lowest level tasks, process flows, and decision points. Figure 5 provides the IMPRINT task network of the baseline model.

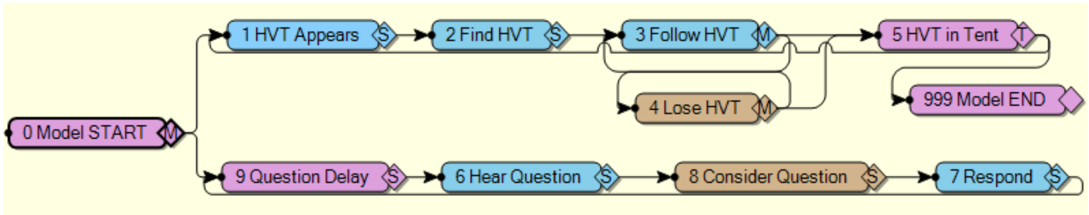


Figure 5: IMPRINT Task Network of Scenario 4

After the baseline was created, tasks were added to represent new automation tasks and new interaction between the human and automation. Table 5 details how the automation was represented using the different stages and levels of automation. The bolded words in the table represent the distinct actions that make each of the levels and stages different from each other. More information on the description of each automation combination can be found in Appendix A.

Table 5: Description of each automation combination

		Levels			
		Three	Five	Seven	Ten
Stages	Information Acquisition	**Automation suggests** three different search patterns for the **human to select**. This is represented in the model by displaying different search pattern suggestions using a pop-up window.	**Automation selects** an alternative search patternand requests confirmation from the human to use the search pattern. The **human approves** or denys the search pattern. If denied, the process is repeated.	**Automation selects and approves** an alternative search pattern and **informs human** of search pattern chosen. It is represented by displaying the chosen search pattern in a pop-up window.	Automation choses an alternative. The **automation completes the task** by executing the search pattern immediately (no window).
	Decision and Action Selection	**Automation suggests** HVT by highlighting every person in the virtual environment with a green color. All potential targets are highlighted in a red color (only in sufficient zoom level). The **human selects** a HVT, and the other highlights are removed.	When the HVT is on the screen, a**utomation selects** and highlights the HVT with a green color (only in sufficient zoom level). The automation requests confirmation via pop-up window. The **human approves** the request and the highlight turns from green to red.	When the HVT is on the screen, **automation selects and approves** the HVT with a red color and **informs human** of the HVT selection via pop-up window. The human then follows the target.	When the HVT is on the screen, a**utomation completes the task** by highlighting the HVT in red (no window). Human then follows red HVT.
	Action Implementation	Once HVT is located by human, **automation suggests** that the target be clicked via pop-up window. The **human selects** the HVT, and then the automation takes over control of the camera and follows the HVT.	Once HVT is located by human, **automation selects** and highlights a specific target on the screen and requests confirmation via pop-up window. The **human approves** or denys the target. If denied, process is repeated.	Once HVT is located by human, **automation selects and approves** a specific target and **informs human** that the target will be followed via a pop-up window. The automation then follows the HVT.	Once HVT is located by human, **automation completes the task** by highlighting and following the target (no window).

Generating IMPRINT Workload and Performance Values

Each model within IMPRINT was set to the same starting number in a random number seed (RNS), originally chosen to be 11, and ran to replicate each trial 300 times. As a result, each of the thirteen models generated an output of 300 total performance values, corresponding to 1200 HVT appearances as 4 HVTs appeared during each trial. Because IMPRINT only records workload values for the first replicate, a macro was applied to run 47 additional replications in which the RNS was incremented from 11-58 and the resulting 48 average workload values were recorded.

As the same RNS were used to initiate each of the models, the data from each of the models was paired, permitting a paired t-test to be applied to compare the baseline model to the alternative models.

Automation Assumptions

It is assumed that each of the distributions applied in the model are an accurate representation of the participant pool. It is also assumed that each automation implementation is accurately represented in the automated models. The primary action (searching and following the target) and the secondary action (answering a mathematics question) are completed in parallel, assuming that the subjects focused on both of these actions at the same time. The communication score is not included in the analysis because the secondary task is unaffected by the automation implementations. The system tasks added in to the automated models are assumed to take no amount of time while the human tasks added into the automated models are assumed to follow micromodels in IMPRINT. The micromodels used for each task can be found in Appendix A along with the descriptions of the respective automation implementations. A full list of the assumptions listed by model task node can be found in Appendix B.

Model Validation

To validate the IMPRINT baseline model, performance data and VACP values for workload were gathered as outputs from the model. Performance values were compared between the subject performance scores and the model scores for Scenario 4 using a t-test with an alpha of 0.05. The p-value for the t-test was 0.323, thus concluding that there is no statistically significant difference between the model scores and the experiment scores, which is the desired result for satisfactory validation. Figure 6 and Figure 7 show the distributions of the primary performance scores.

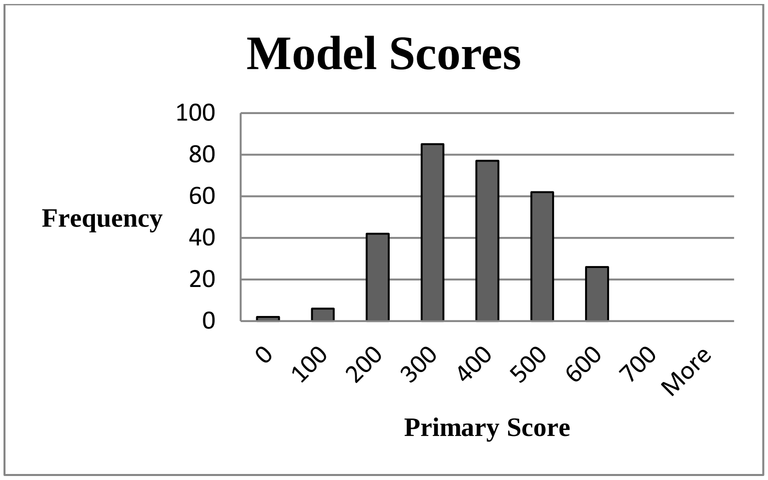


Figure 6: Histogram of the baseline model performance

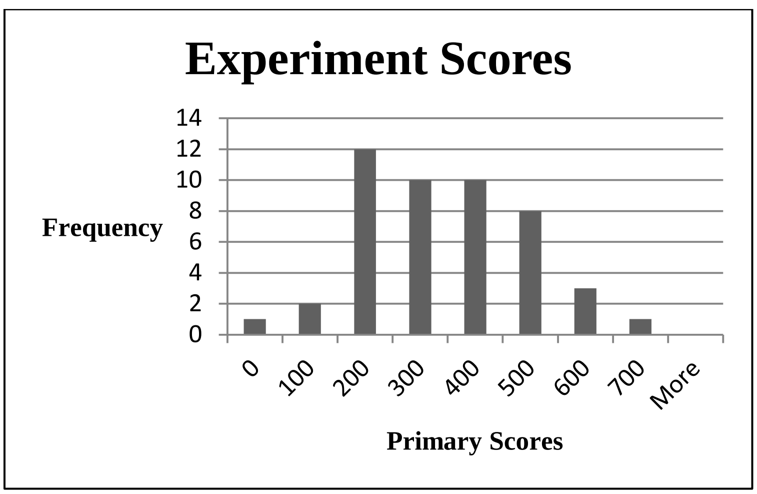


Figure 7: Histogram of the experiment performance

NASA-TLX values were gathered from the human subject experiment, so a comparison was necessary to use the VACP values that IMPRINT works with. Because NASA-TLX and VACP use different scales, t-tests are not feasible for validation of the model VACP values. Instead, an Analysis of Variance (ANOVA) was used to validate the workload scores between the NASA-TLX and VACP values. All four scenarios were used to identify any relationship between the scenarios. If there was a relationship between the scenarios for the human experiment, then the models would be expected to reflect a similar relationship. For example, in the top ANOVA, Scenarios 1 and 3 show very little difference. The bottom ANOVA should then reflect that same relationship, also showing little difference between Scenarios 1 and 2. For the VACP value, a time-weighted average was computed to provide a single value for each of the trials. Figure 8 illustrates a One-way ANOVA between the NASA-TLX score and the Scenario and between the VACP Time Persistent Average and the Scenario.

```
One-way ANOVA: VACP Time Persistent Average versus Scenario

Source     DF       SS      MS      F      P
Scenario    3   2.2819  0.7606  18.44  0.000
Error     188   7.7533  0.0412
Total     191  10.0352

S = 0.2031   R-Sq = 22.74%   R-Sq(adj) = 21.51%


                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level   N    Mean  StDev  --+---------+---------+---------+-------
1      48  14.541  0.217  (-----*-----)
2      48  14.743  0.216                     (----*-----)
3      48  14.554  0.176    (----*-----)
4      48  14.783  0.200                         (----*-----)
                          --+---------+---------+---------+-------
                          14.50     14.60     14.70     14.80

Pooled StDev = 0.203
```

```
One-way ANOVA: TLX Score versus Scenario

Source     DF     SS   MS     F      P
Scenario    3   1500  500  2.10  0.102
Error     188  44851  239
Total     191  46351

S = 15.45   R-Sq = 3.24%   R-Sq(adj) = 1.69%


                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level   N   Mean  StDev  ---------+---------+---------+---------+
1      48  37.99  16.69     (----------*----------)
2      48  42.95  16.22               (----------*----------)
3      48  36.82  14.14  (----------*----------)
4      48  42.92  14.58               (----------*----------)
                         ---------+---------+---------+---------+
                              36.0      40.0      44.0      48.0

Pooled StDev = 15.45
```

Figure 8: ANOVA of VACP and TLX Score vs the Scenario

As shown, both the VACP score and the NASA TLX score follow the same pattern showing that Scenarios 1 and 3 are lower in workload while Scenarios 2 and 4 are higher in workload with little difference between Scenarios 1 and 3 and between Scenarios 2 and 4. While the pattern indicates the same tendencies, none of the differences in the NASA-TLX are statistically significant, due to the large variability between subjects in reporting NASA-TLX scores.

Results and Discussion

Hypothesis 1: All of the automated models will have statistically significant improved performance from the baseline.

The first hypothesis stated that all of the automation models would have statistically significant improved performance values over the baseline system. This hypothesis was partially supported because nine of the twelve models had statistically significant improved performance, shown in Table 6.

Table 6: T-Test Performance Difference in Means (100% Reliability–Baseline)

	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage (A)	72.9**	59.32**	70.9**	65.5**
Decision Stage (C)	90.8**	221.67**	222.59**	231.75**
Response Stage (D)	7.99	9.3	11.9	24.39*

Legend: **p-value<=0.01; *p<=0.05; Grayed out=not significant

Three of the four performance values in the response stage were not statistically different from Baseline. Therefore, it would appear that in the current scenario automation implemented in the action implementation stage has little effect upon performance. This is an unexpected result because it shows how little the automation increased system performance in the stage where automation is traditionally implemented. Thus, the operator performed the action of following the target relatively well. In this instance of automation, the automation did not aid in the process of finding the target. Because the human still had to find the target manually, there was no change to that portion of the task. Once found, the automation would take over and while it never lost the target, the human lost the target infrequently in manual mode (baseline scenario) so there were very little performance points to be gained by automating this stage of the task.

The four performance values in the information acquisition stage were higher than the Baseline, providing a statistically significant difference between all of the information acquisition stage models and the Baseline. This was an expected result. Since the automation is helping the operator find the target by taking control of the camera movement and implementing search patterns, the operator should find the target in less time, resulting in a better score.

The four performance values in the decision stage were higher than the Baseline, providing a statistically significant difference between all of the decision stage models and the Baseline. This was also an expected result, but surprisingly the result is much higher than automation in the information acquisition stage, with the exception of Level 3 Decision Stage. The models predict that the three higher level decision stages will experience a 65 percent increase in performance over the baseline, higher than the 20 percent increase of the highest-scoring information acqisition stage. The higher LOA three Decision Stage models have significantly higher performance than any other automation implementation.

Hypothesis 2: Each of the stages will have statistically different performance from one another.

The second hypothesis stated that each of the stages will have statistically different performance from one another. This hypothesis was supported, with statistical differences between each of the stages, shown in Figure 9. Note that Level 3 Decision

Stage is very similar to a few of the information acquisition stage models. Illustrated in Figure 10 is a Tukey Test confirming the same hypothesis that the stages are different from each other, as none of the intervals in any of the tests contain 0.

```
One-way ANOVA: Primary Score versus Automation

Source        DF        SS       MS       F      P
Automation    12  27332592  2277716  146.17  0.000
Error       3887  60569715    15583
Total       3899  87902308

S = 124.8   R-Sq = 31.09%   R-Sq(adj) = 30.88%


                                     Individual 95% CIs For Mean Based on
                                     Pooled StDev
Level                N   Mean  StDev  -----+---------+---------+---------+----
Baseline           300  340.0  127.8   (-*-)
Level 10 Stage A   300  405.5  134.5            (-*-)
Level 10 Stage C   300  571.7   98.9                                   (-*-)
Level 10 Stage D   300  364.4  145.9      (-*-)
Level 3 Stage A    300  412.8  137.7             (-*-)
Level 3 Stage C    300  430.9   98.7                (-*-)
Level 3 Stage D    300  332.0  127.9  (-*-)
Level 5 Stage A    300  399.3  146.1           (-*-)
Level 5 Stage C    300  561.6  101.1                                 (-*-)
Level 5 Stage D    300  349.3  127.5     (-*-)
Level 7 Stage A    300  410.8  131.4             (-*-)
Level 7 Stage C    300  562.6  102.4                                 (-*-)
Level 7 Stage D    300  351.9  128.4     (-*-)
                                      -----+---------+---------+---------+----
                                          350       420       490       560

Pooled StDev = 124.8
```

Figure 9: ANOVA of Performance Scores vs Automation Implementation

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Stage

Individual confidence level = 98.07%


Stage = Stage A subtracted from:

Stage     Lower  Center  Upper  -------+---------+---------+---------+--
Stage C  111.6   124.0  136.4                           (*)
Stage D  -83.9   -71.5  -59.0            (*)
                                -------+---------+---------+---------+--
                                    -120         0       120       240


Stage = Stage C subtracted from:

Stage      Lower  Center   Upper  -------+---------+---------+---------+--
Stage D  -207.9  -195.4  -183.0  (*)
                                  -------+---------+---------+---------+--
                                      -120         0       120       240
```

Figure 10: Tukey Tests Comparing Stages (Performance)

Hypothesis 3: As the level of automation increases, the performance will also increase.

The third hypothesis stated that the performance would increased as the level of automation increased. The analysis partially supports this hypothesis, with 3 of the 6 comparisons finding differences between levels and 1 of the 6 finding marginal difference as shown in Figure 11.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Level

Individual confidence level = 98.97%


Level = Level 10 subtracted from:

Level     Lower  Center  Upper  -+---------+---------+---------+--------
Level 3  -73.9   -55.6  -37.2   (----*----)
Level 5  -37.9   -19.5   -1.2             (----*-----)
Level 7  -32.2   -13.9    4.5               (----*----)
                                -+---------+---------+---------+--------
                               -70       -35         0        35


Level = Level 3 subtracted from:

Level     Lower  Center  Upper  -+---------+---------+---------+--------
Level 5   17.7    36.0   54.4                             (----*-----)
Level 7   23.4    41.7   60.1                               (----*----)
                                -+---------+---------+---------+--------
                               -70       -35         0        35


Level = Level 5 subtracted from:

Level     Lower  Center  Upper  -+---------+---------+---------+--------
Level 7  -12.7     5.7   24.0                   (-----*----)
                                -+---------+---------+---------+--------
                               -70       -35         0        35
```

Figure 11: Tukey Tests comparing Levels (Performance)

This result stands out because of the impact the different levels made within a particular stage of automation. The levels were hypothesized to provide as much change to the model as the stages did, but some comparisons show no difference, as opposed to the stages which showed significance in all of the comparisons. The only level that was statistically different from all of the others was Level 3. Level 3 did not contain 0 within the interval, thus showing statistical difference between Level 3 and the other three levels. Level 10 also statisitcally differs from both Level 3 and Level 5. Thus the Levels on the extremes (3 and 10) produce more differences than those in the middle (Level 5 and 7).

Performance Results Discussion

Given that the reputation of automation assisting an operator with an RPA task is favorable, the results produced by the response stage are surprising. Automation is generally believed to help accomplish a task better and faster, so no change in the performance is unexpected. However, given the specific automation implementation used, little change in the performance is understandable. The specific action performed by the automation in the action stage is an action widely used by current RPA systems. The automation becomes much more beneficial when used over a period of hours because humans are worse at monitoring a video feed than the automation over extended durations. The human study may not have subjected the operators to trials long enough for this automation advanatage to have been fully realized.

The information acquisition stage results are more consistent with the belief that automation is useful. They provide moderate improvement to a task that the operator was performing, adding a beneficial increase in performance.

The decision stage also represents automation that is not used frequently in an RPA system. Much of the choice is left up to the operators when categorizing individuals who have appeared on a video feed. Designers may struggle with a proper solution that can differentiate between people and choose one that fits a certain description, but if it were possible to build such automation, it may provide considerable benefit to the operators.

Hypothesis 4: All of the automated models will have statistically significant reduced workload from the baseline.

The fourth hypothesis stated that all of the automated models would have workload changes that were statistically lower than the baseline. This hypothesis was supported by the difference in means paired t-tests shown in Table 7. The information acquisition and decision stage models were significant, but magnitude of the change was largely irrelevant compared to the response stage. When incorporating automation into the RPA task, one of the goals was to reduce the operator workload. Illustrated in Table 7 are the workload results comparing the baseline model with no automation to the twelve automation models. There are a few unexpected results with regards to the workload.

Table 7: T-Test Workload Difference in Means (Automation–Baseline)

	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage (A)	-0.1859**	-0.1980**	-0.1709**	-0.1642**
Decision Stage (C)	-0.1367**	-0.6256**	-0.6740**	-0.3832**
Response Stage (D)	-2.951**	-2.380**	-2.476**	-2.494**

Legend: **p-value<=0.01; *p<=0.05;

The response stage has the most noticable workload reduction. Every level in the response stage had a greatly reduced workload when compared to the baseline and even the rest of the automated models. Table 7 shows how great the difference becomes, with greater than a 2 point reduction in workload. The mean time-weighted average workload for the baseline model is 14.78, thus the increase shown by each response stage model is approximately a 15% or greater increase over the baseline model. This reduction is three times as much as any of the other automated models with the next largest reduction, at Level 7 Decision Stage, reporting a 5% increase over the baseline. The reason for this stems from the action being completed by the automation. In the response stage, the automation completes the task of following the target, reducing the operator's task to a monitoring task, which requires much less workload than the act of continuously recentering the camera video feed.

Automating the information acquisition stage does not produce a large change in workload. This is a surprising result considering that this automation also removes the action of recentering the screen. Although the t-test results show that the information acquisition stage models are all significant when compared to the baseline, they still represent the smallest workload change from the baseline out of all of the models.

Automating the decision stage consisted of a moderate change in workload, generally a greater reduction than the information acquisition stage, but less of a change than the response stage. This is not too surprising, given how the automation was implemented for the decision stage. The operator continued most of the tasks similar to the baseline, but the automation would attempt to locate the target along with the operator. The automation may have allowed for speed of identification, but the responsiblity of identification was still held by the operator, thus workload was minimally affected by the automation.

Hypothesis 5: Each of the stages will have statistically different operator workload from one another.

The fifth hypothesis stated that each of the stages would have statistically different operator workload from one another. This hypothesis is supported and illustrated in the ANOVA provided in Figure 12 and the corresponding Tukey Tests provided in Figure 13, where the four response stage models can be seen on the left side of the graph and the other models can be seen on the right side of the graph in the ANOVA. The Tukey Tests show how the information acquisiton stage and the decision stage are similar, but still significant because the intervals do not contain the value 0.

One-way ANOVA: VACP Values versus Automation

```
Source       DF       SS      MS       F      P
Automation   12  728.761  60.730  159.91  0.000
Error       611  232.041   0.380
Total       623  960.803

S = 0.6163   R-Sq = 75.85%   R-Sq(adj) = 75.37%


                                     Individual 95% CIs For Mean Based on
                                     Pooled StDev
Level              N    Mean  StDev  ---+---------+---------+---------+------
Baseline          48  14.783  0.200                               (-*-)
Level 10 Stage A  48  14.619  0.210                             (-*-)
Level 10 Stage C  48  14.400  0.160                           (-*-)
Level 10 Stage D  48  12.290  1.234      (-*-)
Level 3 Stage A   48  14.597  0.198                             (-*-)
Level 3 Stage C   48  14.647  0.150                              (*-)
Level 3 Stage D   48  11.833  0.992  (*-)
Level 5 Stage A   48  14.585  0.227                             (-*-)
Level 5 Stage C   48  14.158  0.143                         (-*)
Level 5 Stage D   48  12.403  1.029       (-*-)
Level 7 Stage A   48  14.612  0.191                             (-*-)
Level 7 Stage C   48  14.109  0.146                        (-*-)
Level 7 Stage D   48  12.307  1.034      (-*-)
                                     ---+---------+---------+---------+------
                                     12.0      13.0      14.0      15.0


Pooled StDev = 0.616
```

Figure 12: ANOVA of Baseline Workload Scores vs Automation

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Stage

Individual confidence level = 98.04%


Stage = Stage A subtracted from:

Stage       Lower   Center    Upper  -------+---------+---------+---------+--
Stage C  -0.4324  -0.2751  -0.1179                (*)
Stage D  -2.5527  -2.3954  -2.2382  (*)
                                     -------+---------+---------+---------+--
                                         -1.5       0.0       1.5       3.0


Stage = Stage C subtracted from:

Stage       Lower   Center    Upper  -------+---------+---------+---------+--
Stage D  -2.2776  -2.1203  -1.9630    (*)
                                     -------+---------+---------+---------+--
                                         -1.5       0.0       1.5       3.0
```

Figure 13: Tukey Tests comparing Stages (Workload)

With two stages so close together, the designer should consider the small change in workload when deciding between the information acquisition stage and the decision stage; however the response stage has significantly reduced workload when compared to either of the two stages or the baseline and should first be considered for feasibility before the other two stages.

Hypothesis 6: As the level of automation increases, the workload will decrease.

The sixth hypothesis stated that as the levels of automation increased, the workload would decrease. This hypothesis was not supported by the analysis, as both Figure 14 and Table 7 show that the levels had a very small impact, if any, on the difference in workload.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Level

Individual confidence level = 98.95%


Level = Level 10 subtracted from:

Level      Lower  Center  Upper  ---------+---------+---------+---------+
Level 3  -0.458  -0.077  0.304   (--------------*--------------)
Level 5  -0.435  -0.054  0.327    (--------------*--------------)
Level 7  -0.474  -0.093  0.288  (--------------*---------------)
                                 ---------+---------+---------+---------+
                                       -0.25      0.00      0.25      0.50


Level = Level 3 subtracted from:

Level      Lower  Center  Upper  ---------+---------+---------+---------+
Level 5  -0.358   0.023  0.404        (--------------*--------------)
Level 7  -0.397  -0.016  0.365      (--------------*---------------)
                                 ---------+---------+---------+---------+
                                       -0.25      0.00      0.25      0.50


Level = Level 5 subtracted from:

Level      Lower  Center  Upper  ---------+---------+---------+---------+
Level 7  -0.420  -0.039  0.342    (--------------*---------------)
                                 ---------+---------+---------+---------+
                                       -0.25      0.00      0.25      0.50
```

Figure 14: Tukey Tests comparing Levels (Workload)

Workload Results Discussion

The response stage automation is a type of automation that is currently being used in a variety of RPAs, albeit in a different context. Most of the monitoring that an operator completes is related to the flight of the aircraft. Designers have become adept at incorporating automation designed to fly the RPA and while this does reduce the workload substantially, designers need to be careful not to underload the operator. In a situation where the operator does not have any tasks to complete, situation awareness drops and boredom can set in.

The information acquisition is a stage of automation that may not currently be used frequently when incorporating automation into RPAs. However, this result indicates that system desginers may not want to focus on automating any sensor movement, as the operators performed similarly to the automation when in charge of the sensors. Also, the sensor portion of the task is not what makes up most of the workload during that time. Most of the workload is due to the operator performing the visual search task in an attempt to find the HVT. So even when the automation is able to remove a portion of the workload, that portion was not large enough to result in a substantial decrease in workload.

The models in the decision stage are an example of automation that increases the performance dramatically while leaving the workload relatively unchanged. The significance between the baseline model and the automated models still indicates statistical significance, but the magnitude of the change is relatively limited when compared to results from the response stage. This type of automation would be very helpful to desingers that felt the operator workload level was comfortable, but wanted to increase the performance of the system. Designers also need to keep in mind the nonlinear relationship between workload and performance when making automation design decisions.

Conclusions

This paper shows how workload and performance can be affected by different implementations of automation. Stages and levels of automation were used to create different combinations of automation, which were then incorporated into an RPA task.

The levels within a stage produced slight variation with regards to the primary task performance, but different stages affected the performance to a greater extent. The information acquisition stage provided a moderate increase in the performance, the decision stage provided a large increase in the performance, and the response stage provided no discernable increase in performance. The performance did not change as a result of decreased operator workload or increased performance in the primary task. Automation reduced the operator workload for all of the automated models. The information acquisition stage and decision stage models saw a small decrease in workload. The response stage provided a large decrease in comparison to the other automated models. The change in workload due to changes in levels of the automation was indiscernable.

The largest increase in performance occurred for all of the decision stage models because the automation was reducing the time it took to find the target. Based off of the results, the actual decision making took the longest time for the human to complete, leaving a large amount of time for the automation to reduce, adding many points to the performance score. With regards to the workload, the response stage models greatly reduced the amount of workload that the operator experienced. The automation allowed the cognitive workload of the operator to reduce from a following task to a simpler monitoring task. A reduction in workload may be small, but the small decrease grows as the time following the target increases. Automation can be invaluable when attempting to assist the operator or the system. However, in order to obtain the best results from the automation implementation, system designers will need to understand how different implementations may affect the system.

Future Work

Future work in this area includes further examination of the relationship between the stages and levels to discern which combinations work together optimally. Performing this same investigation with other systems will aid in discovering if the preferred stage-level combination differs from system to system or is common across systems. If some combinations work better than others in all systems, this would greatly aid in reducing the design trade-space.

While these results provide an insight into using different automation for RPA operations, future research should focus on implementing these stages and levels combinations of automation into a human subject study. Some effects may not be noticed in DES that a human study may uncover.

When making automation implementation tradeoffs, other factors, such as reliablity may also impact operator workload and system performance. Future work should seek to identify these factors and examine their impacts with on workload and performance with regards to the different combinations of stages and levels of automation. If one combination has less sensitivity than another, it may be prudent to choose the less sensitive combination

IV. The Impact of Reliability on the Performance and Operator Workload Within a System

Abstract

This paper investigates how automation reliability may affect the workload and performance of the operator as well as how the impact of reliability is affected by the different automation implementations. This study uses IMPRINT discrete event simulation to evaluate three levels of reliability in twelve different baseline automation implementations. The automation implementations incorporate different instances of automation into a remotely piloted vehicle task by varying the stage and level of automation. The reliability is assessed at 100%, 80%, 70%, and 60%. The results indicate that the performance values between 100% reliability and reduced reliability are generally significantly reduced with the exception of the response stage models. The results for the workload values indicate very little change between 100% reliability and the reduced reliability. The performance between the baseline models and the reduced reliability models experiences some significant changes while the workload between the baseline models and the reduced reliability models is insensitive to change.

Introduction

Understanding Reliability

As defined by Parasuraman et al., automation "refers to the full or partial replacement of a function previously carried out by the human operator" (Parasuraman, Sheridan, & Wickens, 2000). Incorporating automation into industrialized systems brought with it new changes to the way systems were designed. By adding automation, systems became more complex and more robust, creating a paradox in which the more complex the system is, the more crucial a human will be to keeping the system running properly (Bainbridge, 1983). A complex system can also be helpful for completing difficult tasks but incorporating automation can be difficult due to the complexity. A complex system has a higher potential for error because of how many more areas a problem can arise from. More parts mean more places the system can fail.

The goal of incorporating automation in a system is to minimize errors (usually attributed to the human), but not every error-causing situation can be foreseen by the designer. The more errors within the automation, the worse the automation will perform. At some reliability level, the automation will begin to start degrading the performance of the system. The point at which the degradation begins differs based on the automation implementation chosen. Some implementations may have less sensitivity to reliability, allowing those implementations to outperform the others. This research aims to aid system designers in choosing the most effective automation implementation given the degraded reliability.

Reliability and RPAs

Reliability of a system becomes extremely important if there is minimal human contact to intervene in the systems operations. Space missions where a probe was sent out into the solar system to collect data on another planet required parts to be far more reliable than a machine in a production line with a human standing next to it to make sure the job gets done properly. Frequently, when automation fails, human intervention is necessary (Bainbridge, 1983). If no human can reach the system, then the failure may never be fixed. In the case of remotely piloted aircraft (RPA), a machine that is flying without a human in the cockpit, the direct human contact will be minimal compared with manned systems. RPAs have human pilots flying the aircraft but if a failure occurs, the geographically separated operator may be unable to recover the aircraft before it crashes. Any RPA conducting reconnaissance may contain sensitive information about the enemy. Because of the cost implications associated with RPA crashes, reliability of the parts and reliability of the automation continues to receive attention (Dixon, Wickens, & Chang, 2005).

As the complexity increases in a system, the automation may need to accept more tasks to keep the human from becoming overworked. As the automation receives more tasks from the human, the human must be aware of possible errors and ways to fix them. If the automation is unable to execute the tasks properly, then the human may be required to intervene in order to correct the automation. In some instances of faulty automation, the overall system performance may be better off without the automation. Gauging the point at which the automation becomes harmful may be difficult without any previous data gathered about the automation to know when or how it fails.

Research Goals

This paper investigates the relationship between the automation and its reliability in terms of how those factors affect operator workload and system performance. In addition to examining reliability, this study also examines the interaction between reliability and different types of automation implementation. The study uses discrete-event simulation (DES) to model a human subject experiment for RPA operations. The DES model of the baseline systems is expanded to incorporate 12 different automation implementations. Each implementation is then examined on three levels of reliability in order to determine how automation failures impact operator workload and the overall system performance.

Background

RPAs and Workload

Current RPA missions rely upon multiple operators to control a single aircraft. In a time where the military is reducing the workforce, the number of operators needs to be reduced. One of the limiting factors on the operator is the amount of cognitive workload that can be handled at one time. Reducing that workload requires automation. Automation supports the operator by assuming control of some of the tasks, reducing the stress on the operator workload. However, much of the automation incorporated currently is not perfect. There is a potential that for a portion of time, the automation will act sub-optimally, causing a decrease in the mission performance that otherwise would not have occurred had the third operator remained. The likelihood of sub-par mission performance can be reduced with better information about how automation should be implemented into the system and information about any secondary effects that are not immediately visible to the designer.

Automation

Automation is contained within many of the tasks we perform in a day. Daily tasks on a computer use automation constantly so the human does not have to become overburdened with simple tasks. In that sense, the human is able to focus on the pressing issues that are more worthwhile. However, automation may not always support the operator. If the automation fails or the automation cannot communicate properly with the operator, the automation may prevent the operator from effectively accomplishing the task. Any harmful interference from the automation could add to operator workload rather than reduce it.

In addition to potentially making the task more difficult for the operator, automation may create new actions for the operator to complete. In most cases, these actions do not require as much cognitive workload as the task the automation is performing, but typically the automation does not completely remove a task from the task load of the operator. For example, most automated tasks require some form of interaction between the automation and the operator. If the automation provides notifications about a system failure, the human must still react to that notification. The human does not completely shed the task, but requires less workload than when working with a system with no automation. The automation is still considered to be effective because it reduced the overall workload on the operator. In cases where the operator is overloaded and performance is degraded, adding automation can reduce the risk of potential failures.

Automation provides some unique advantages and disadvantages. One advantage is a general reduction in human error. By moving human interaction with the system into a monitoring position, the human participation in the task is reduced (Swanson, et al., 2012). With the human slightly removed from the task, the accompanying human error is normally lessened. Also, when the automation is incorporated correctly, the overall task load of the operator will be reduced. By reducing the human’s task load, the human operator is able to focus on other tasks that may improve overall system performance.

One of the disadvantages of automation is that reducing human participation will likely result in reduced operator situation awareness. If automation takes over key processes and the human lacks the appropriate situation awareness, then the human may be unable to effectively resolve automation failures. Furthermore, reduced interaction with the system can lead to a degradation of operator's skillsets. Conversely, more interaction with the task increases the operator's skill level and better prepares them to make decisions in unexpected situations.

Automation can also potentially cause an increase in workload because of the added communication between the system and the operator. Examples of this additional communication include: asking the operator to choose the task to complete, asking for permission to begin the task, informing the operator that it is beginning a new task, asking the operator to select between multiple courses of action, and notifying the operator of task status/completion.

As mentioned above, a reduction in human error is expected when automation is implemented. Clumsy implementation of automation may, however, lead to an increase in human error (Woods, Johannesen, Cook, & Sarter, 1994). New burdens may be unintentionally placed on the operator, creating more problems and more opportunities for error, along with the expected benefits provided by the automation (Woods, Johannesen, Cook, & Sarter, 1994). For example, if automation is only built to accommodate routine scenarios, then latent problems may arise when a scenario appears that was not covered. These latent problems could then emerge when the human works through the scenario (Woods, Johannesen, Cook, & Sarter, 1994). That scenario may never occur, but the possibility of it happening leads to an added possibility of human error due to the clumsy implementation of automation.

Stages and Levels of Automation

As automation replaces tasks performed by the human operator, replacement may include tasks related to any of the four stages of human information processing: Sensory Processing, Perception/Working Memory, Decision Making, and Response Selection. Sensory Processing gathers information from the outside world and provides it for higher level processing. Perception/Working Memory synthesizes this information with remembered information to form an interpretation of the environment. Decision Making relies upon the interpretation of the environment to decide upon a course of action. Response Selection completes the action decided upon in the Decision Making stage. When automated, the replacement technologies are referred to as Information Acquisition, Information Analysis, Decision and Action Selection and Action Implementation, respectively. The corresponding stages for machine information processing are shown in Figure 15 (Parasuraman, Sheridan, & Wickens, 2000).

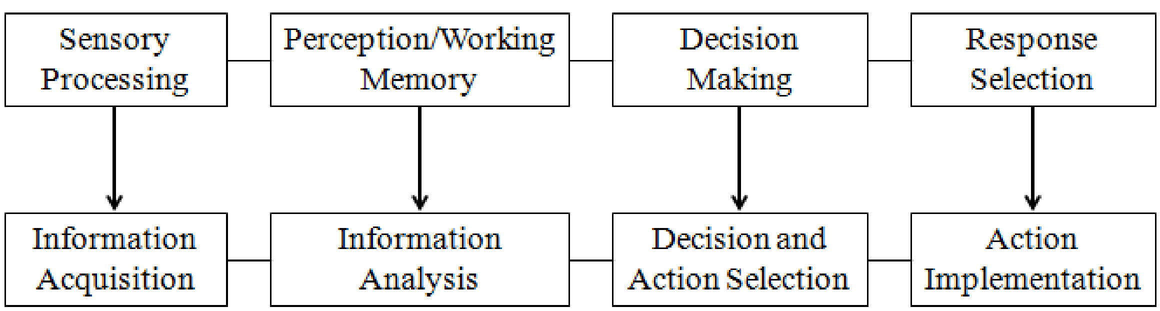


Figure 15: Stages of machine processing built from the human information processing model – adapted from (Parasuraman, Sheridan, & Wickens, 2000)

The replacement technology can automate each of the four stages of information processing to one of ten levels of automation, as proposed by Sheridan and Verplank (1978). These ten levels of automation (LOAs) are provided in Table 8. Combined, the stages and levels form forty combinations of automation that are unique from each other. For example, an Information Acquisition stage coupled with level three will produce automation that gives several different choices on how information should be obtained. If the level was changed from three to five, then the automation may only ask the human if the choice chosen by the automation should be used or not. Conversely, if the stage was changed from Information Acquisition to Decision and Action Selection but remained at level three, then the automation may ask the operator to choose from a set of actions to complete. The combination of stages and levels of automation provides numerous design options for implementing automation into a system.

Table 8: Levels of Automation – adapted from (Sheridan & Verplank, 1978)

	Determines Alternatives	Suggests Alternative	Selects Alternative	Executes Alternative	Informs of Action
Level 1	Human	Human	Human	Human	N/A
Level 2	Computer	Human	Human	Human	N/A
Level 3	Computer	Computer	Human	Human	N/A
Level 4	Computer	Computer	Computer, Human may or may not approve	Human	N/A
Level 5	Computer	Computer	Computer	Computer, if Human approves	N/A
Level 6	Computer	Computer	Computer	Computer, unless Human vetoes	N/A
Level 7	Computer	Computer	Computer	Computer	Always
Level 8	Computer	Computer	Computer	Computer	If Human requests
Level 9	Computer	Computer	Computer	Computer	If Computer decides to inform human
Level 10	Computer	Computer	Computer	Computer	N/A

Reliability

Reliability is also partly a function of system complexity. As systems become more complex, the automation becomes more complex as well, leaving greater opportunities for unforeseen problems that could lead to a system failure. This results in the “irony of automation” where, as the complexity of a system rises, human involvement becomes more critical due to all of the unforeseen problems (Bainbridge, 1983).

Recent reliability studies in the RPA field focus on the reliance and compliance of human dependence (Wickens & Dixon, 2006). Reliance is the state of human dependence when the automation is quiet. Compliance is the state of human dependence when the automation is alerting the human that something has potentially gone wrong. Human reliance stays high when the automation has fewer misses, meaning that the human has more trust that the system is fine when the automation is quiet. Human compliance stays high when the automation produces fewer false alarms, meaning that the human has more trust in the automation to correctly identify when something has gone wrong. When both metrics are high, the human experiences less cognitive workload because the human believes that the automation is handling the task well. Both of these metrics are based on human perception, so there is potential for a disconnect between actual automation performance and perceived automation performance. A study performed by Dixon and Wickens (2006) illustrates the reliance and compliance of the human and how those two metrics may affect the reaction time of the human to any automation signals. Dixon and Wickens found that when the automation produced more misses, the operator was quicker to notice them and fix them, but had trouble completing the concurrent tasks in a timely manner (less reliance). When the automation produced more false alarms, the operator had a slower and less accurate response (less compliance) to the alarm but showed little change in the ability to complete the concurrent tasks.

Reliance and compliance are important attributes for alarm-style automation systems; however, these attributes may be less relevant for other types of automation implementation. For example, with RPA operations, the automation may help track a target. This example does not fit in neatly with reliance and compliance which are geared towards alerts and alarms, thus reliance and compliance may be less helpful in determining the reliability of the automation. Another way to look at reliability is the percentage of time that the automation does not fail, represented as a number from 0-100% (Parasuraman, Molloy, & Singh, 1993). A failure can represent any type of action taken by the automation that the operator did not expect or any type of halt in the automation sequence, where it cannot manage to complete assigned activities. Previous automation studies have attempted to identify the point at which automation failure makes the system performance decrease and operator workload increase above the baseline of not having any automation at all. One study has placed this number at approximately 70-75% reliability (Wickens & Dixon, 2006). Thus, if the automation fails more than 25-30% of the time, then the operator would have performed better without the automation. However, the task being completed also has an impact on the effectiveness of the automation as the reliability is reduced. John and Manes found that even automation reliabilities below 70% still may be helpful (John & Manes, 2002). In their study, the goal of the operator was to locate a target while the automation would provide suggestions on places to look. As the reliability was reduced below 70%, the automation was still helpful in aiding the operator. Thus, the reliability threshold for which it begins to harm the workload and performance of the operator may depend on the task being completed. Perhaps metrics including task completion times for the human and the automation, recovery time necessary in the event of a reliability failure, and operator workload could be useful in further understanding this tradeoff. System designers need to know at what threshold the automation reliability should stay above in order to help, rather than hinder, task performance.

Discrete Event Simulation and IMPRINT

In order to capture the reliability of the automation, this study uses discrete event simulation (DES) to model the workload and performance of an operator completing a common RPA task. Simulations provide several advantages over human experiments including a decrease in the amount of time to run trials, less outside factors to influence the subjects (i.e. recent family death, loss of job), and the ability to evaluate multiple manipulations of the system. A sample amount of information is necessary to build a simulation, but given that information, many different types of manipulations can then be accomplished. The simulation is constructed using the Improved Performance Research Integration Tool (IMPRINT), a DES environment specifically tailored to model human performance (Alion Science and Technology, 2009). IMPRINT enables the quantitative modeling of operator workload through incorporation of the Visual, Auditory, Cognitive, and Psychomotor (VACP) scale. VACP draws on the multiple resource workload theory to quantitatively assign demand to resource channels using verbal descriptions of categories of tasks. There are seven channels within the VACP model: the visual, auditory, cognitive, fine motor, gross motor, tactile, and speech. As a task is completed, the operator experiences varying levels of workload in each of these channels which combine to form a single unique value for overall workload. Originally developed for US Army acquisitions, IMPRINT can be used to assist in the research of human performance (Alion Science and Technology, 2009).

Purpose

This paper demonstrates the impact of reliability levels on operator workload and system performance. This research extends previous reliability studies by examining automation reliability across the spectrum of automation stages and levels. Identifying the interactions between reliability and automation implementation will enable system designers to make more effective tradeoffs when incorporating automation.

To evaluate the impact of reliability and automation implementations, this research identifies and answers eight hypotheses. The eight hypotheses can be broken down into two sets of four. The first set consists of four hypotheses that are related to the system performance and the second set consists of four hypotheses that are related to the operator workload. Both sets assess the same independent variables, with the first hypothesis addressing the difference between the lower reliability models and the baseline model with no automation, the second hypothesis addressing the difference between the different reduced reliability models and their respective 100% reliability model, the third hypothesis addressing the difference between the automation stages at each reliability measure, and the fourth hypothesis addressing the difference between the automation levels at each reliability measure. All eight hypotheses are as follows:

Set 1 (System Performance Hypotheses)

1) All of the models at 60% reliability will have significantly reduced performance when compared to the baseline with no automation.

2) All of the models at 80%, 70%, and 60% will have significantly reduced performance when compared to their respective 100% model.
3) The performance differences between stages will be significantly affected by changes in the reliability measures.
4) The performance differences between levels will be significantly affected by changes in the reliability measures.

Set 2 (Operator Workload Hypotheses)

5) All of the models at 60% reliability and above will have significantly reduced workload when compared to the baseline with no automation.
6) All of the models at 80%, 70%, and 60% will have significantly increased workload when compared to their respective 100% model.
7) The workload differences between stages will be significantly affected by changes in the reliability measures.
8) The workload differences between levels will be significantly affected by changes in the reliability measures.

Methodology

Human RPA Experiment

The RPA task consists of a surveillance operation where the goal is to locate a high value target (HVT) within a marketplace, shown in Figure 16. Once the operator had located the HVT, designated by a rifle held in both hands, the operator would notify the system that the HVT was found, and would then track the HVT until the HVT left the screen. Each trial consisted of following 4 HVTs, all of which appeared sequentially, so only one HVT was visible at a time. The operator had the task of controlling the sensor feed in order to find the HVT. Performance points were awarded for tracking the HVT upon acknowledgement that the target had been found.

Figure 16: Screenshot of market in Surveillance Task

In addition to the primary task, the operator also had to complete a secondary communication task designed to represent communication with other pilots or air traffic controllers. The communication task consisted of a mathematics question related to the RPA’s altitude or airspeed, which was provided both over audibly over a headset and in text for on the right-most screen, as shown in Figure 17.

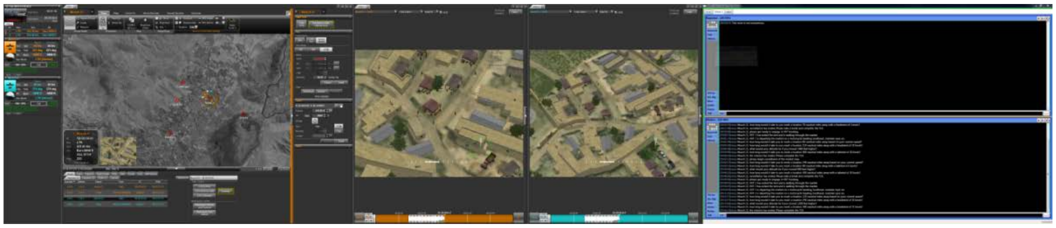


Figure 17: Complete setup of displays in human experiment

The surveillance task consisted of four different scenarios intended to vary the difficulty of the primary task. The four scenarios combined two independent variables, the amount of distractors (high or low) and the camera quality (high or low). For evaluating reliability and automation implementation, this research focuses on the most difficult scenario with high distractors and low camera quality because this scenario is the most suitable candidate for incorporating automation.

Baseline Model

This paper builds upon previous work from Chapter III. Modeling the Effects of Stages and Levels of Automation on Operator Workload and System Performance in RPA Operations. The previous work developed a baseline simulation in IMPRINT that modeled the performance and workload of a human operator conducting an RPA surveillance task. This simulation model used performance and behavior data from a human-in-the-loop study conducted by the 711th Human Performance Wing at Wright Patterson AFB, OH to determine the task network, decision logic, and probabilistic task times. See Methodology in Chapter III for a detailed description of the baseline model. From this baseline model, twelve automation combinations out of the possible forty (4 stages x 10 levels of automation) were modeled to evaluate how different automation implementations impacts operator workload and system performance (see Experimental Design for DES Automation Experiment).

Model Validation

To validate the IMPRINT baseline model built from the human experiment, performance data and VACP values for workload were gathered as outputs from the model. Performance values were compared between the subject performance scores and the model scores for Scenario 4 using a t-test with an alpha of 0.05. The p-value for the t-test was 0.32, thus finding no statistical difference between the model scores and the experiment scores. An Analysis of Variance (ANOVA) was used in order to validate the workload scores. To compare the NASA-TLX and VACP values, a time-weighted average was found for the VACP values. The single value of the VACP average and NASA-TLX was then compared across all of the trials and was found to have no statistical significance. For more information on the model validation, refer to Model Validation in Chapter III.

Generating IMPRINT Workload and Performance Values

Each model within IMPRINT was set to the same starting number in a random number seed (RNS), originally chosen to be 11, and ran to replicate each trial 300 times. As a result, each of the thirteen models generated an output of 300 total performance values, corresponding to 1200 HVT appearances as 4 HVTs appeared during each trial. Because IMPRINT only records workload values for the first replicate, a macro was applied to run 47 additional replications in which the RNS was incremented from 11-58 and the resulting 48 average workload values were recorded.

As the same RNS were used to initiate each of the models, the data from each of the models was paired, permitting a paired t-test to be applied to compare the baseline model to the alternative models.

Automation Assumptions

It is assumed that each of the distributions applied in the model are an accurate representation of the participant pool. It is also assumed that each automation implementation is accurately represented in the automated models. The primary action (searching and following the target) and the secondary action (answering a mathematics question) are completed in parallel, assuming that the subjects focused on both of these actions at the same time. With regards to the communication task, it is assumed that the automation implementations will have no effect on the secondary task, so the secondary communication task is not included in the analysis. The system tasks added in to the automated models are assumed to take no amount of time while the human tasks added into the automated models are assumed to follow micromodels in IMPRINT. The micromodels used for each task can be found in Appendix A along with the descriptions of the respective automation implementations. A full list of the assumptions listed by model task node can be found in Appendix B.

Reliability Assumptions

It is assumed that the automated models are a valid representation of the automation actions portrayed. In addition, it is assumed that the reliability failure occurring in each of the models can be immediately reset by the operator. Upon reset, the reliability will once again have a chance of failure. The human is also assumed to have no loss in faith when the automation fails, so no matter how many times the automation fails the human will continue to operate the same way. With respect to failures, the human is memoryless. It is also assumed that any failure in automation will not disrupt any other portion of the system besides the current portion the automation is working within. A deeper look into the assumptions with regards to the reliability can be found in Appendix B.

Experimental Design for DES Automation Experiment

After baseline model creation and validation, twelve alternative models were created to model the implementation of automation. Out of the forty possible combinations (4 stages x 10 levels of automation), the twelve combinations selected enable a significant reduction in the number of alternatives to analyze while still spanning the entire design space for the automation. The three selected stages are: Information Acquisition (Stage A or information acquisition stage), Decision and Action Selection (Stage C or decision stage), and Action Implementation (Stage D or response stage). Out of the four stages, the information analysis stage (Stage B), was not chosen because the information analysis stage was very similar to the information acquisition stage for the RPA task. Any changes that affected the acquisition stage would also affect the information analysis stage. The four levels are levels three, five, seven, and ten. Note that Level 1 automation represents the original baseline model. Each of the automation actions was applied to the baseline automation. Table 9 provides descriptions of the different levels and stages that were used in each of the twelve models.

Table 9: Descriptions of Automation Actions

		Levels			
		Three	Five	Seven	Ten
Stages	Information Acquisition	**Automation suggests** three different search patterns for the **human to select**. This is represented in the model by displaying different search pattern suggestions using a pop-up window.	**Automation selects** an alternative search patternand requests confirmation from the human to use the search pattern. The **human approves** or denys the search pattern. If denied, the process is repeated.	**Automation selects and approves** an alternative search pattern and **informs human** of search pattern chosen. It is represented by displaying the chosen search pattern in a pop-up window.	Automation choses an alternative. The **automation completes the task** by executing the search pattern immediately (no window).
	Decision and Action Selection	**Automation suggests** HVT by highlighting every person in the virtual environment with a green color. All potential targets are highlighted in a red color (only in sufficient zoom level). The **human selects** a HVT, and the other highlights are removed.	When the HVT is on the screen, a**utomation selects** and highlights the HVT with a green color (only in sufficient zoom level). The automation requests confirmation via pop-up window. The **human approves** the request and the highlight turns from green to red.	When the HVT is on the screen, **automation selects and approves** the HVT with a red color and **informs human** of the HVT selection via pop-up window. The human then follows the target.	When the HVT is on the screen, a**utomation completes the task** by highlighting the HVT in red (no window). Human then follows red HVT.
	Action Implementation	Once HVT is located by human, **automation suggests** that the target be clicked via pop-up window. The **human selects** the HVT, and then the automation takes over control of the camera and follows the HVT.	Once HVT is located by human, **automation selects** and highlights a specific target on the screen and requests confirmation via pop-up window. The **human approves** or denys the target. If denied, process is repeated.	Once HVT is located by human, **automation selects and approves** a specific target and **informs human** that the target will be followed via a pop-up window. The automation then follows the HVT.	Once HVT is located by human, **automation completes the task** by highlighting and following the target (no window).

First, each automation combination was run at 100% reliability. Although it is helpful to understand how the automation changed the performance and the workload of each operator, the reliability of the automation will never be 100%. Past research showed that automation that has failed 25-30% of the time (70-75% reliable) tends to degrade the task performance and raise the operator workload. Because of this, a potential error was created for each of the twelve automation combinations to understand how a failure might affect operator workload and system performance. Table 10 provides a description of each of the failures.

Table 10: Description of Reliability Failures

		Levels			
		Three	Five	Seven	Ten
Stages	Information Acquisition	Failure occurs when search pattern only **covers a certain percentage of the market**. This is represented with the Automation Pass/Fail task. The human may realize there is a problem and restart the automation. In that case, **automation suggests** new search patterns and the **human selects** one of the suggestions.	Failure occurs when search pattern only **covers a certain percentage of the market**. This is represented with the Automation Pass/Fail task. The human may realize there is a problem and restart the automation. In that case, **automation selects** a new search pattern and the **human approves** the suggestion.	Failure occurs when search pattern only **covers a certain percentage of the market**. This is represented with the Automation Pass/Fail task. The human may realize there is a problem and restart the automation. In that case, the **automation selects and approves** a new pattern and the **human is informed** of the selection.	Failure occurs when search pattern only **covers a certain percentage of the market**. This is represented with the Automation Pass/Fail task. The human may realize there is a problem and restart the automation. The **automation completes the task** again to choose a new pattern.
	Decision and Action Selection	Failure occurs when the automation **does not highlight the potential HVTs or highlights a distractor**. The human may realize there is a problem and restart the automation. In that case, the **automation suggests** new potential HVTs and the **human selects** one of the suggestions.	Failure occurs when the automation **does not highlight the potential HVT or highlights a distractor**. The human may realize there is a problem and restart the automation. In that case, **automation selects** a new potential HVT and the **human approves** the suggestion.	Failure occurs when the automation **does not highlight the potential HVT or highlights a distractor**. The human may realize there is a problem and restart the automation. In that case, the **automation selects and approves** a new HVT and the **human is informed** of the selection.	Failure occurs when the automation **does not highlight the potential HVT or highlights a distractor**. The human may realize there is a problem and restart the automation. In that case, the **automation completes the task** again to choose a new HVT.
	Action Implementation	Failure occurs when the automation **begins to follow a distractor or nothing at all**. In that case, the human may skip the notification that a target was lost and restart the automation. The human must then relocate the target, at which point the **automation suggests** new HVTs to follow and **human selects** one of the suggestions.	Failure occurs when the automation **begins to follow a distractor or nothing at all**. In that case, the human may skip the notification that a target was lost and restart the automation. The human must then relocate the target, at which point the **automation selects** a new HVT to follow and **human approves** the suggestion.	Failure occurs when the automation **begins to follow a distractor or nothing at all**. In that case, the human may skip the notification that a target was lost and restart the automation. The human must then relocate the target, at which point the **automation suggests and approves** a new HVT to follow and the **human is informed** of the selection.	Failure occurs when the automation **begins to follow a distractor or nothing at all**. In that case, the human may skip the notification that a target was lost and restart the automation. The human must then relocate the target, at which point the **automation completes the task again** to follow a new HVT.

Independent and Dependent Variables

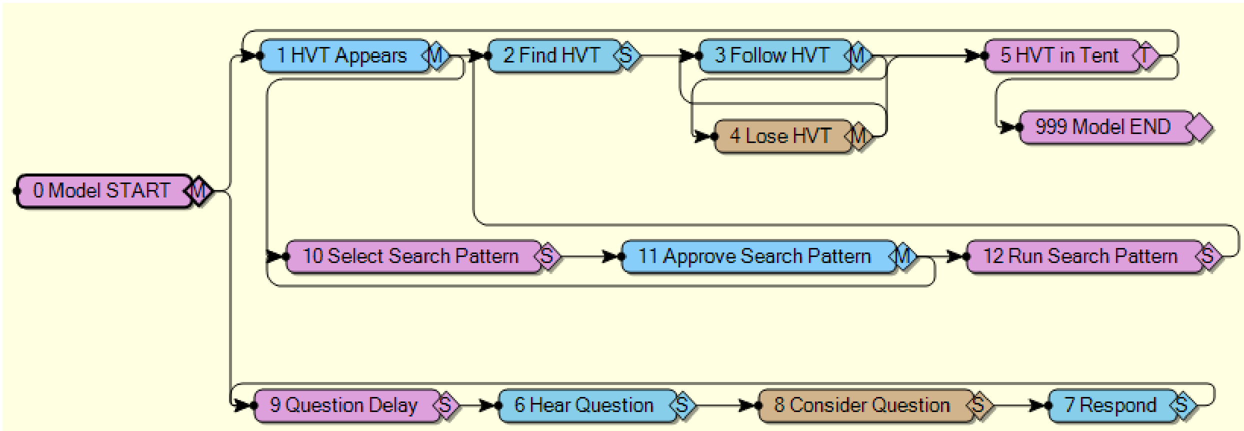
This research evaluates two independent variables: automation implementation and degree of reliability. Automation implementation consists of the 12 stage and level combinations: the Information Acquisition stage, the Decision and Action Selection stage, and the Action Implementation stage with level three, five, seven, and ten. The degree of reliability altered the likelihood that an automation error would occur. For example, if the likelihood was 80% reliability, then the automation error would only happen for 20% of the automated task occurrences. Each automation implementation contains a task with a probability of failure and the probability is assessed each time the task is performed. Depending on the outcome, the model will continue down either the success or failure path, reevaluating a failure every time the task is performed. Note that because the task will repeat, there is potential for multiple failures to occur in a single task run. The three degrees of reliability used in each of the combinations were 80%, 70%, and 60%. Thus, the experimental design consisted of 12x4 = 48 alternative designs (12 automation implementations, 4 degrees of reliability) to compare to the original 12 baseline automation implementations at 100% reliability.

There were two dependent variables within this DES. The first one was the performance of the operator, and was based out of a total scored of 1000 points. Every time the operator would designate that the target was found with the F key, the operator would start accumulating points at a rate of one point every third of a second. That accounted for 800 of the total points. The other 200 came from the mathematics question, where 50 points would be given for a right answer, -5 for a wrong answer, and 0 for no answer. The primary performance values in the baseline model averaged out to

340 points. The second dependent variable was the workload of the operator which was the time-weighted average VACP values gathered from the IMPRINT models. The VACP values were added up over the whole trial period and then divided by the amount of seconds within the trial to gather the time-weighted average. The time-weighted workload values in the baseline model averaged out to a score of 14.78. The communication score is not included in the analysis because the secondary task is unaffected by the automation implementations.

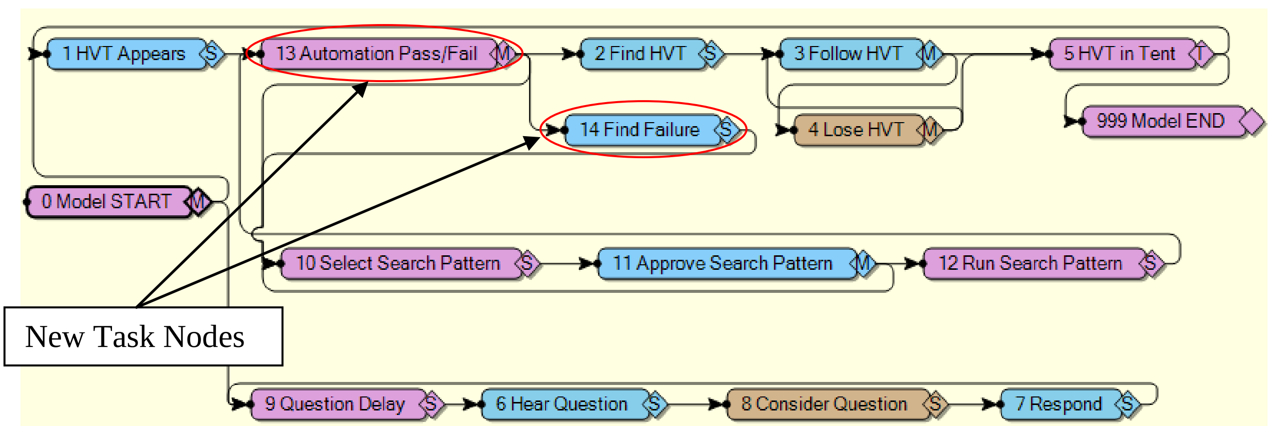
Implementing Reliability into the Automation Implementation Models

Each automation implementation model needed to be modified to account for the consequence of the potential failure caused by the degraded reliability. For example, the automation action of the Level 5 Stage A model was to select a search pattern and request approval from the operator to use the selected pattern. When the reliability was 100%, the automation performed as intended. When the reliability was reduced to 70%, additional nodes were required to determine whether or not a failure occurred, and to capture the alternative tasks caused by the failure. In the case of 70% reliability, the automation would fail 30% of the time that the automated task occurred and when a failure occurred, only a portion of the market was searched. This partial search would be unsuccessful in finding the target, and the process would begin again with the selection of the search pattern after the partial search was conducted. Figure 18 and Figure 19 show the model at 100% reliability and again at 70% reliability within IMPRINT.



Legend: Purple – system task, Blue – task containing workload, Brown – task containing workload with no performance gain

Figure 18: Level 5 Stage A (information acquisition stage) at 100% reliability



Legend: Purple – system task, Blue – task containing workload, Brown – task containing workload with no performance gain

Figure 19: Level 5 Stage A (information acquisition stage) at 70% reliability

Similar tasks were added to each of the twelve automation models to capture the probability and consequence of failure, resulting in forty-eight new models (4 levels of reliability for each of the 12).

An Analysis of Variance (ANOVA) was used in order to evaluate the workload and performance of the models. The ANOVA provided a 95% confidence interval of the performance and workload values for each of the models. Coupled with that, a paired t-test, with a significance level of 0.05, was used to evaluate the difference in means between the 100% reliability models and the degraded reliability models.

Results and Discussion

Hypothesis 1: All of the models at 60% reliability will have significantly reduced performance when compared to the baseline with no automation.

The first hypothesis stated that all of the models at 60% reliability will have significantly reduced performance when compared to the baseline with no automation, shown in Table 11. This hypothesis was partially supported by the results. The negative values in the table represent instances where the model at 60% reliability performed worse than the baseline model while the positive values in the table represent the times where the 60% reliability instance performed better than the baseline model. The response stage models only had three implementations that were significantly lower when compared to the baseline and the information acquisition stage models had one implementation that was significantly lower. Thus, the performance in the information acquisition stage models at 60% reliability was very similar to performance with no automation at all. All decision stage models show significantly improved performance, illustrating the improvement in system performance even with reduced reliability.

Table 11: T-Test Performance Difference in Means (60% Reliability–Baseline)

	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage (A)	-18.5	-18.62*	-1.3	6.4
Decision Stage (C)	19.36*	131.6**	132.96**	165.63**
Response Stage (D)	-34.12**	-21.1*	-18.43*	11.23

Legend: *p-value<=0.05 **p<=0.01 Grayed out=not significant

This result is unexpected given the information from the past studies. As one study pointed out, once automation degrades below 70-75% reliability, the system performs worse with automation than with no automation at all. This result illustrates that the degrading of the reliability may be dependent upon the stage of automation.

Hypothesis 2: All of the models at 80%, 70%, and 60% will have significantly reduced performance when compared to their respective 100% model.

The second hypothesis stated that all of the models at 80%, 70%, and 60% would have significantly reduced performance when compared to their respective 100% reliability models, shown in Table 12. This hypothesis was largely supported, with only four implementations producing values that are not deemed significant. Table 12 shows the results of the paired t-tests for the performance scores between the baseline reliability of 100% and the other reliabilities of 80%, 70%, and 60% for each automation combination. The table values provide the difference in means for the corresponding paired t-test. To obtain the difference in means, the lower reliability performance score was subtracted from the baseline of 100%. Therefore, a negative value indicates that the model with the lower reliability had the lower performance score as well. A positive value indicated that the lower reliability had a higher performance score. To determine whether the p-value was statistically significant, an alpha of 0.05 was used; asterisks are used in the table to capture the level of significance. Most of the models resulted in significantly lower performance even at the higher 80% reliability model, showing how volatile the performance scores are when reliability changes.

Table 12: T-Test Performance Difference in Means (X Reliability–100% Reliability)

X = 80% Reliability				
	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage	-54.1**	-25.1*	-35.1**	-32.3**
Decision Stage	-46.1**	-62.9**	-59.3**	-40.3**
Response Stage	-19.3*	2	0.9	-24.0*
X = 70% Reliability				
	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage	-70.8**	-56.6**	-59.8**	-49.8**
Decision Stage	-59.3**	-82.3**	-81.0**	-54.2**
Response Stage	-29.2**	-5.4	-8	-31.6**
X = 60% Reliability				
	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage	-106.2**	-95.2**	-88.6**	-74.8**
Decision Stage	-86.5**	-104.9**	-104.5**	-90.0**
Response Stage	-41.0**	-16.4*	-19.1**	-28.0**

Legend: *p-value<=0.05 **p<=0.01 Grayed out=not significant

Levels 5 and 7 of the response stage show significance only when comparing 100% reliability to 60% reliability. A high increase in performance due to the benefits of perfect automation would be expected to result in a high decrease in performance as the automation reliability decreases, but Levels 3 and 5 of the response stage show resistance to change.

Hypothesis 3: The performance differences between stages will be significantly affected by changes in the reliability measures.

The third hypothesis stated the performance differences between stages will be significantly affected by changes in the reliability measures. This hypothesis was supported and illustrated in Figure 20. The interaction p-value is below 0.05, meaning there is a significant interaction between the stage factor and the reliability factor. This means that the difference in performance between stages changes as the reliability changes. The two factors influence each other so that the amount of change in performance values from one stage to another depends upon the reliability measure.

Two-way ANOVA: Primary Performance versus Stage, Reliability

Source	DF	SS	MS	F	P
Stage	2	62768738	31384369	1960.32	0.000
Reliability	3	9533416	3177805	198.49	0.000
Interaction	6	1949750	324958	20.30	0.000
Error	14388	230349738	16010		
Total	14399	304601641			

S = 126.5   R-Sq = 24.38%   R-Sq(adj) = 24.32%

Figure 20: 2-Way ANOVA comparing Performance Values of Different Stages and Reliabilities

Hypothesis 4: The performance differences between levels will be significantly affected by changes in the reliability measures.

The fourth hypothesis stated that the performance differences between levels will be significantly affected by changes in the reliability measures. This hypothesis was not supported, and can be seen in Figure 21. The interaction p-value is 0.84, which is much higher than the significance threshold of 0.05, thus the level of reliability does not impact the difference caused by as change in level. This means that there is very little, if any, interaction between the reliability measures and the levels of automation, so a change in one of the factors will consistently result in the same change across the instances of the other factor.

Two-way ANOVA: Primary Performance versus Level, Reliability

Source	DF	SS	MS	F	P
Level	3	7865485	2621828	131.35	0.000
Reliability	3	9533416	3177805	159.21	0.000
Interaction	9	98680	10964	0.55	0.839
Error	14384	287104061	19960		
Total	14399	304601641			

S = 141.3 R-Sq = 5.74% R-Sq(adj) = 5.65%

Figure 21: 2-Way ANOVA comparing Performance Values of Different Levels and Reliabilities

Performance Results Discussion

As expected and shown in Table 13, decreased reliability produced lower performance scores, as can be seen with all of the statistically significant differences in means reporting a negative score. For 80% and 70% reliability in Level 5 Action Implementation stage (response stage) and Level 7 Action Implementation stage, the numbers are not statistically significant. In other words, these combinations for each of the three reduced reliabilities produced performance scores that were not statistically different from the baseline of 100% reliability. Furthermore, 60% reliability for Level 5 and 7 with the response stage represented the smallest difference in means for each of their respective levels. For every level of automation, regardless of how badly the automation performed, the change in reliability had the least effect on the response stage.

This could be for a number of reasons, but one of the more probable ones is that at the response stage, the automation is only performing the function of following the target. The automation in the response stage has no effect on how quickly the HVT can be found, so the performance score is not affected by the automation during what is believed to be the major contribution to the performance. System designers, if designing a system with automation to increase the performance, may want to identify stages that affect the system performance and incorporate automation into those stages.

It can be noted that Level 5 Decision Stage contains all of the highest difference in means besides Level 3 Information Acquisition Stage at 60% reliability. These differences are a reduction of about one quarter of the entire primary task score from 100% reliability to 60% reliability. These differences generated p-values below 0.05, thus they are statistically significant. In other words, all of the performance scores differ greatly between Level 5 Decision Stage with 100% reliability and Level 5 Decision Stage with less-than-100% reliability. In this case, every drop in reliability results in a performance drop. Although Level 3 Information Acquisition Stage had the highest difference in means as a single model with regards to performance, all twelve decision stage models regardless of the level and reliability had high differences. These differences illustrate how much of an effect there was because of the change in reliability. In general, the decision stage requires a lot of time to complete. In other systems, most of the time may be spent in other stages such as the response stage, but when a system requires the operator to continually make small decisions, the decision stage becomes one of the primary stages that the operator spends most of the time.

In addition, the interaction between the performance values of the reliability measures and the stages and levels produced some interesting results. The interaction between the reliability and the stages resulted in significance, meaning that a change in instance of one of the factors will affect the differences between the levels of the other factor. For example, the performance values of the decision stage and the response stage may grow closer or further apart as they change with reliability changes. The interaction between the reliability and the levels produced insignificant results, thus a change in one of the factors does not affect differences between levels of the other factor.

Hypothesis 5: All of the models at 60% reliability and above will have significantly reduced workload when compared to the baseline with no automation

The fifth hypothesis stated that all of the models at 60% reliability and above will have significantly reduced workload when compared to the baseline with no automation. This hypothesis was largely supported, as nine of the twelve models showed significance when compared to the baseline shown in Table 13. Also to note, all four of the response stage models continued to show significantly reduced workload at a low reliability level. This illustrates that even as the reliability starts to decrease, the workload is generally significantly lower when the automation is incorporated than not.

Table 13: T-Test Performance Difference in Means (60% Reliability–Baseline)

	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage (A)	-0.0425	-.1625**	-.1218**	-.0240
Decision Stage (C)	-.0338	-.3637*	-.4169**	-1.165**
Response Stage (D)	-2.536**	-2.237**	-2.201**	-2.318**

Legend: *p-value<=0.05 **p<=0.01 Grayed out=not significant

Hypothesis 6: All of the models at 80%, 70%, and 60% will have significantly increased workload when compared to their respective 100% model.

The sixth hypothesis stated that all of the models at 80%, 70%, and 60% reliability will have an increased workload when compared to their respective 100% reliability models. This hypothesis was partially supported, showing significance in about half of the models and no significance in the other half, shown in Table 14. Within the table, all of the values represent the workload value at 100% reliability subtracted from the workload value at the reduced reliability. Any value that is positive shows an increased workload as reliability is reduced while any value that is negative shows a decreased workload as reliability is reduced. Also to note, the models at 80% reliability show significance in half of the models, illustrating how even a smaller reduction in reliability can significantly affect the workload of the operator.

Table 14: T-Test Workload Difference in Means (X Reliability–100% Reliability)

X = 80% Reliability				
	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage	.1055*	0.0046	0.0153	.0880*
Decision Stage	0.0464	.2175**	.2033**	-0.2544**
Response Stage	.255*	-0.01	0.089	0.047
X = 70% Reliability				
	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage	0.124**	0.027	0.009	0.104*
Decision Stage	0.053	0.542**	0.24**	-0.379**
Response Stage	0.252	0.129	0.272*	0.039
X = 60% Reliability				
	Level 3	Level 5	Level 7	Level 10
Information Acquisition Stage	0.143**	0.036	0.049	0.14**
Decision Stage	0.103**	0.262**	0.257**	-0.781**
Response Stage	0.415*	0.143	0.276	0.176

Legend: *p-value<=0.05 **p<=0.01 Grayed out=no significance

Ten of the twelve decision stage models show significance, so reliability seems to have an effect on workload; however some implementations show increasing workload as the reliability decreases and some implementation show decreasing workload as the reliability decreases. This result is unexpected, but considering how much workload can be devoted to a decision, greater workload changes may occur. Designers may want to keep in mind the fact that the operator workload in the decision stage is reliant upon the reliability of the automation.

Hypothesis 7: The workload differences between stages will be significantly affected by changes in the reliability measures.

The seventh hypothesis stated that the workload differences between stages will be significantly affected by changes in the reliability measures. This hypothesis was not supported, shown in Figure 22. The interaction p-value is 0.086, which is above the threshold of 0.05, thus failing to reject the null hypothesis of no interaction. This means that as one of the factors changes, the other factor will change the same across all of the levels of that factor: when comparing the change due to reliability of two different stages, the change will be consistent across the levels. This result was unexpected, as the interaction between the stages and reliability measures when comparing performance values was significant.

Two-way ANOVA: Workload Values versus Stage, Reliability

Source	DF	SS	MS	F	P
Stage	2	2444.95	1222.47	2496.29	0.000
Reliability	3	4.70	1.57	3.20	0.023
Interaction	6	5.44	0.91	1.85	0.086
Error	2292	1122.43	0.49		
Total	2303	3577.50			

S = 0.6998   R-Sq = 68.63%   R-Sq(adj) = 68.47%

Figure 22: 2-Way ANOVA comparing Workload Values of Different Stages and Reliabilities

Hypothesis 8: The workload differences between levels will be significantly affected by changes in the reliability measures.

The eighth hypothesis stated that the workload differences between levels will be significantly affected by changes in the reliability measures. The findings do not support this hypothesis, with Figure 23 showing how automation levels within the same stage continued to change at similar rates as reliability changed. This result is illustrated through the interaction p-value, which produced a value of 0.795, much higher than the threshold for significance of 0.05. This result implies that changing the automation levels does not have much of an effect on the differences between the reliability measures, and vice versa.

Two-way ANOVA: Workload Values versus Level, Reliability

Source	DF	SS	MS	F	P
Level	3	5.73	1.90840	1.23	0.298
Reliability	3	4.70	1.56503	1.01	0.389
Interaction	9	8.45	0.93833	0.60	0.795
Error	2288	3558.64	1.55535		
Total	2303	3577.50			

S = 1.247   R-Sq = 0.53%   R-Sq(adj) = 0.00%

Figure 23: 2-Way ANOVA comparing Workload Values of Different Levels and Reliabilities

Workload Results Discussion

Table 14 shows the results of the paired t-tests for the workload values between the baseline reliability of 100% and the other reliabilities of 80%, 70%, and 60% for each automation combination. The table values provide a difference in means between 100% reliability and either 80%, 70%, or 60% reliability for each automation combination. To determine whether the p-value was statistically significant, an alpha of 0.05 was used; asterisks are used in the table to capture the level of significance.

One of the few takeaways from this table is that the values in the decision stage levels are mostly significant. Excepting Level 3 Decision Stage at 80% and 70%

reliability, every other decision stage had a high statistical significance. That significance illustrates how the workload differed between the baseline reliability and the two alternative reliabilities. This could be attributed to how much the action in the decision stage influenced the overall performance and workload. Most of the workload and performance changes that the operator experienced were attributed to deciding upon an HVT, so the automation should have the largest effect when taking on that role.

Another unexpected result can be seen when looking at all of the response stage levels in Table 14. Six of the eight differences between means are not statistically significant when using an alpha of 0.05. In other words, there is a low likelihood that there is a difference between the workload of the operator when automation is following the target with 100% reliability and 80%, 70%, or 60% reliability. Even if the reliability drops to levels below the threshold that the automation is helping, the operator does not see any significant workload change. This is important because it shows how little of an effect the reliability has on the task. If a designer chooses to implement automation for a similar task, then the designer may not want to spend the extra money to bring the reliability above 90% if it does not provide any benefits for the operator.

Just like the performance, these results indicate how much of an impact the reliability of the automation had on the operator workload. As Table 14 illustrates, the automation implementation has a large effect on the workload. Much of the significance is dependent upon the stage and level of automation. From 70% to 60% reliability, three of the twelve models experienced a change from significance to non-significance or vice versa. While some change occurred based on the reliability, most of the change seemed focused around the automation that was used. This research did assume that the operators would act in the same manner regardless if the automation was 100% or 60%, so some of the results may change if operator reaction is involved. Judging by the results, if designers decide to incorporate automation into some of the key decision making tasks, precautions need to be taken in order to improve the reliability and keep the operator workload reduced.

One more observation focuses on the values in Table 13. This table illustrates the differences between the 60% reliability models and the baseline models with no automation. Most of the values in the table are still significantly negative, suggesting that even when the reliability of the automation drops to 60%, the operator still feels less workload than when the system is using no automation. The three models that are not significant are Level 3 Decision Stage and Levels 3 and 10 Information Acquisition Stage. This table illustrates how helpful automation may be, even with a reduction in reliability. This result largely contradicts previous research suggesting that automation should only be used when reliability is above 70-75% reliable.

Finally, the last two hypotheses produced unexpected results, showing no significance for the interaction between the stages of automation and the reliability measures and no significance for the interaction between the levels of automation and the reliability measures. These results mean that when the reliability is reduced from 70% to 60%, the difference between workload values within a level 3 model and level 10 model of the same stage are the same. These results are unexpected because higher automation levels would expect to see larger differences between the workload values as reliability is reduced.

Conclusion

Key Findings

These results indicate how important the automation implementation and the reliability are to the success of the system. The different types of implementation affect both the performance and the workload of the operator, some implementations more than others. Not a single stage and level was superior in every way, so designers will need to consider different choices depending on their needs. If a system is performing well but the operator is consistently overworked, then automation may need some type of monitoring task to reduce that workload. If both the system is performing poorly and the operator is overworked, then it may be possible that more than one implementation is necessary. As automation becomes more necessary to use for more complex systems, designers will need to understand what the operator needs and how the automation interacts with the operator.

Furthermore, based off of the results from the performance scores and the workload values, the area where the automation brought about the most change was during the actual decision selection. When the automation took over much of the decision making process, the human had the greatest reduction in workload and the largest change in performance. Based off of these results, if the designer was to implement automation, a stage that may result in improved performance and reduced workload is during the decision selection phase.

Following the same idea, the designer must have a high reliability for the automation when the automation performs well, or the high gains will be reduced by high losses. This can also hold true for any system. If the designer is able to locate the action that presents the greatest effect on the system, or change in the system, then the designer can automate that action and, if done well, can greatly increase the output of the system.

This study finds that automation reliability affects performance and workload differently, with reliability affecting performance but not workload for certain automation implementations, and vice versa. For example if the designer is looking to improve the system performance with automation, then tasks that aid decision making may benefit from automation. If the designer is looking to reduce the amount of workload that the operator experiences, then the system may benefit from automation incorporated at any task that falls under the action implementation stage of the processing model.

Future Work

Future work in automation reliability could focus more on how trust plays a part in how the human accepts the automation. The work presented shows automation reliability as if an operator continued acting in the same manner even when the reliability drops. Trust is a large part of how well the operator and automation function together because if the operator has no trust in the automation, then the operator can never completely hand over the task to the automation. This work illustrates some of the benefits when the operator can completely transfer the task to the automation even in the light of failing automation, but does not take into account how the operator may want to take over for the automation at some point.

Another interesting portion of work that was not addressed in this research focuses on how different implementations may complement one another. If automation was incorporated in multiple stages, the question becomes whether the stages support each other or not. For example, if some automation in the decision stage was implemented and then automation from the response stage followed up on the task the decision stage completed, the handoff of information may be smooth or some information may get lost. Incorporating multiple automation implementations over two or more stages may produce some interesting results. On top of that question, the levels can also play a factor in how much information the automation shares with the operator. Too much automation at a level 10 (fully automatic) may leave the operator with a loss in situation awareness. Unforeseen problems need to be addressed before a system becomes operational or the system will not perform to its fullest potential.

Lastly, the results from this reliability and implementation research can be tested again by human subjects. Because of the differences between DES models and human subjects, using these same automation implementations and reliability measures will expand the knowledge on the reliability and implementation of the automation upon the operator workload and system performance. DES provides a way to quickly run trials and remove some of the variance in human subjects while human subjects can provide real-world data that DES assumed away. Both methods provide unique benefits that, when used together, will make the end results more robust.

V. Conclusions and Recommendations

Chapter Overview

This chapter begins by providing a broad overview of the current situation for remotely piloted aircraft (RPA) in the military. It then restates the research objective posed at the beginning of this paper. The research objective is followed by the two investigative questions and a discussion of their subsequent answers. The chapter then ends with recommendations for future work to extend this research.

Research Motivation

Current trends point towards a steady, increasing growth of the use of RPAs, even in the commercial sector. Recently, Amazon stated in a letter to the Federal Aviation Administration (FAA) that they would like to use RPAs as a way to transport packages in a more timely fashion (Misener, 2014). Within the military, leaders continue to advocate for RPAs, citing the dull, dirty, and dangerous jobs for which RPAs are so well-suited (Van Cleave, 2003). In order to realize the military's future vision, some of the fundamental ways that RPA missions are conducted need to change. Rather than having a one-to-one ratio of human to RPA at best, automation could allow for a single human to control multiple RPAs if designed correctly. With RPAs working as a force multiplier, the military would then be one step closer to reducing manpower while simultaneously increasing effectiveness. This research investigated ways to incorporate increased automation into the RPA system to reduce the workload associated with managing a single RPA. With reduced workload, future operators may be able to control multiple RPAs without becoming overloaded.

Research Objective

The increasing complexity of systems has initiated a need for automation to compliment human efforts to complete the task at hand. Tasks have become more involved due to the desire for increased operator output, thus automation is needed to remove some of the actions from the human when the workload is too high. This research aimed to provide insight to system designers regarding the impact of automation implementation design decisions. A discrete event simulation (DES) was used to simulate operators in a high workload environment in order to determine effective ways to implement automation. The Improved Performance Research Integration Tool (IMPRINT) DES software was used to provide workload and performance data based off of the data gathered from a human experiment completed by the 711th Human Performance Wing.

The experiment centered on humans interfacing with a virtual environment representation of an RPA system. The goal of the study was to locate a HVT within a marketplace. The performance data was based on how long it took the operator to find the target and how well the operator could follow it, while subjective workload data was based on a NASA-TLX questionnaire that the operator completed at the end of each trial. The information gathered was used to build DES models within IMPRINT, which could be modified to change or add tasks, based on the automation portrayed. In IMPRINT, performance was measured in total points awarded using the same mechanism as was done in the human experiment, while workload was measured with VACP values.

Investigative Question One

Two areas are important to investigating design tradeoffs for automation implementation: the stage and level of automation and the reliability of the automation. The first area can be addressed by revisiting the first investigative question identified in Chapter 1:

1. What stages and levels of automation reduce operator workload and increase performance in the surveillance task?

The automation was incorporated into the model as a specific action based off of different stages and levels of automation. The different stages and levels of automation combine to form forty automation implementation combinations. Twelve of these combinations were chosen to be simulated and evaluated. They were deliberately chosen to capture the full range of values to ensuring substantial differences in the implementation of the automation, while also minimizing the number of treatment combinations to be investigated. The stages chosen include the information acquisition stage (acquisition stage or Stage A), the decision and action selection stage (decision stage or Stage C), and the action implementation stage (response stage or Stage D). The levels chosen were levels 3, 5, 7, and 10. Each stage represented a different part of the process that was automated, while the levels represented the amount of automation incorporated. Out of the four stages, the information analysis stage (Stage B), was not chosen because the information analysis stage was very similar to the information acquisition stage for the RPA task. Any changes that affected the acquisition stage would also affect the information analysis stage.

The purpose of creating these twelve models was to develop an understanding of how the baseline performance and workload might compare to the different automated models. The first three hypotheses evaluated the performance dependent variable; hypotheses four through six evaluated the operator workload dependent variable. Each set of three assesses the same independent variables, first addressing the difference between the system with no automation and the system with automation, second addressing the difference between each of the stages of automation, and third addressing the difference between each of the levels of automation.

Performance of Stage and Level Models

The first hypothesis states that all of the automated models would have statistically significant improved performance from the baseline. Four of the models showed no improved performance but eight of the twelve models had statistically significant improved performance. None of the response stage models were significant. This is an unexpected result, but can be explained due to how little the automation affected the performance of the task. The automation did not help the operator find the target, so the time to find the target was relatively the same. The automation followed the target well, but because the operator rarely lost the target, the performance benefit from the automation was minimized.

The second hypothesis states that each of the stages will have statistically different performance from one another. This hypothesis was supported, with statistical differences between each of the stages. All of the decision stage models experienced a large performance increase, the information acquisition stage models experienced a moderate increase, and all of the response stage models experienced a minimal increase over the baseline system. This result illustrates the diverse reaction to the different stages of automation. System designers need to be aware that the stage of automation implementation can have significant impact on system performance outcomes.

The third hypothesis states that the performance increases as the level of automation increases. The analysis did not support this hypothesis and instead, the levels within stages changed very little. This result was unexpected, as increasing the amount of automation for a task is believed to increase the performance as well. System designers should keep in mind that keeping the operator engaged in the task is expected to increase the operator situational awareness.

Workload of Stage and Level Models

The fourth hypothesis states that all of the automated models would have workload changes that are significantly reduced below the baseline. This hypothesis was supported for every model. While some of the stages may not have reduced the workload by large magnitudes of time-averaged VACP values, those small differences can amount to a large reduction in workload when taken over a longer period of time.

The fifth hypothesis states that each of the stages will have statistically different operator workload from one another. This hypothesis was supported by the results. The response stage models had much lower workload than the rest of the models, even though the response stage did not experince substantial increases in performance. The other two stages were closer, but still showed significance between the two stages. This demonstrates that gains in performance and workload are not directly connected, and systems designs need to evaluated for both. In an environment where operator workload is more of a concern than system performance in any system, automation implementation in the response stage could be very useful.

The sixth hypothesis states that as the levels of automation increases, the workload would decrease. This hypothesis was not supported by the data, with differences between levels not producing differences in workload. This result was unexpected because reducing the amount of tasks allocated to the operator would be expected to reduce the amount of workload the operator experiences.

Investigative Question Two

The second area that is important to developing automation in RPAs is the reliability. The reliability can be addressed by revisiting the second investigative question identified in Chapter 1:

2. How does the level of reliability of the automation affect the workload and performance of the user during the task?

After the twelve models were built to model the different automation implementations, each implementation was modified to incorporate three different levels of reliability. The levels were chosen based on previous findings, suggesting that around 70-75% reliability is the point at which the automation harms the operator workload and performance of the system. In order to capture possible patterns outside what was expected, 80% and 60% were included with 70% to create three different levels of reliability. The twelve models from the first investigative question became the baseline models for this portion of the study, representing the automation performing at 100% reliability with no errors. The three reliability models were compared to the respective baseline model that contained the same stage and level of automation to interpret impacts of reliability on workload and performance. The purpose of using the automation models as baseline models and comparing them to models with reduced reliability is to determine how much of an effect the reliability has on the system. A total of eight hypotheses were made to predict the effect of reduced reliability. The eight hypotheses can be divided into two sets of four. The first set consisted of four hypotheses that evaluated the system performance dependent variable and the second set consisted of four hypotheses that evaluated the operator workload dependent variable. Both sets assess the same independent variables, with the first hypothesis addressing the difference between the lower reliability models and the baseline model with no automation, the second hypothesis addressing the difference between the different reduced reliability models and their respective 100% model, the third hypothesis addressing the difference between the automation stages at each reliability measure, and the fourth hypothesis addressing the difference between the automation levels at each reliability measure.

Performance of Reliability

The first hypothesis states that all of the models at 60% reliability will have significantly reduced performance when compared to the baseline with no automation. The information acquisition models did not support this hypothesis, only containing one data point that was significant when compared to the baseline while the decision and response models were generally significant. The results show that the information acquisition models were very similar to the performance with no automation at all, the decision stage models still had significantly better performance values even at 60%

reliability, and the response stage models had significantly worse performance than the baseline.

The second hypothesis states that all of the models at 80%, 70%, and 60% would have significantly reduced performance when compared to their respective 100% reliability models. This hypothesis was largely supported, with only four models producing values that could not be deemed significant (Levels 5 and 7 in the response stage in both the 70% and 80% reliability measures). This result illustrates the effect that the reliability has on the performance of the system and should be taken into consideration by system designers when trying to incorporate automation.

The third hypothesis states that the performance differences between stages will be significantly affected by changes in the reliability measures. This hypothesis was largely supported, with the interaction between the stages and reliability measures showing significance. This means that as reliability changes, the difference between stages of the same level significantly change. This result shows how reliability can affect the performance values of each stage differently.

The fourth hypothesis states that the performance differences between levels will be significantly affected by changes in the reliability measures. This hypothesis was not supported, producing a p-value interaction of 0.84, much higher than 0.05. This result was unexpected because changing reliability measures was expected to change higher level automation more than lower level automation. Instead, the reliability affected both higher and lower level automation in the same manner.

Workload of Reliability

The fifth hypothesis stated that all of the models at 60% reliability will have significantly higher workload when compared to the baseline with no automation. This hypothesis was largely supported, as every level in the response stage showed significance when compared to the baseline and only three of the twelve did not show significance. This result illustrates how insensitive the response stage was to reducing the workload. Even at 60% reliability, the values for the response stage were still much lower than any other stage.

The sixth hypothesis states that all of the models at 80%, 70%, and 60% reliability will have an increased workload when compared to their respective 100% reliability models. This hypothesis was partially supported, showing significance in about half of the models and no significance in the other half. The decision stage models showed significance in all eight models except when comparing 100% reliability to 70% reliability. The information acquisition stage models showed significance in levels 3 and 10. The response stage models showed significance in level 3 at 60% reliability and level 7 at 70% reliability. Note that Level 10 Decision Stage model shows significance in the negative direction, meaning that the 70% and 60% reliability models reported less workload than the 100% reliability model.

The seventh hypothesis states that the models in each stage would have significantly reduced workload as reliability is reduced. This hypothesis was not supported, showing an interaction p-value of 0.08. While close to the threshold of significance of 0.05, this value is still deemed not significant. This hypothesis produced different results than the same hypothesis dealing with the performance values, indicating how the two dependent variables were affected differently by the independent variables.

The eighth hypothesis states that the workload differences between levels will be significantly affected by changes in the reliability measures. The findings do not support this hypothesis, producing results that indicate no interactions between the workload values of the levels and reliability measures. This means that any change in reliability will not significantly affect the differences between reliability levels.

Recommendations for Future Research

While this research focused on stages and levels of automation and reliability, there were areas of reliability that were not covered. Reliance and compliance although researched in previous studies, was not addressed in this paper. Much of the work in developing automation focuses on the human receiving a signal from the automation, informing the human that something is wrong with the plane. This signal-based strategy focuses entirely on the reliance and compliance of the human as the automation signals are perceived by the operator or not. Reliance and compliance may be adapted for use in other automation implementation such as the scenario described in this paper where the operator must search for a target and follow it, instead of in a limited capacity of informing the operator when something is wrong with the plane.

Along the same lines as reliance and compliance, trust is another factor in how well the operator and automation function together within the system. If the operator does not have sufficient trust in the automation, then much of the benefit of the automation could be lost. The operator’s workload remains high because the operator verifies the completion of tasks accomplished by the automation. To increase the complexity of this problem, each operator differs in the amount of trust that is placed in the automation. If the amount of trust can be identified, the amount of automation may be increased or decreased to suit the operator based on both the workload of the operator and the amount of trust the operator has for the automation.

With regards to this experiment, DES provided great flexibility in how different scenarios may be created. However, DES does not provide the same data as real human subjects because of the assumptions that must be made, thus an extension of this research would be to incorporate the different automation implementation found in the DES models into the human experiment to better understand the performance and workload of the operator. Each type of experiment has merits, but each type of experiment also has flaws. Complementing this research with further human subject research would provide greater insight and validity into the findings of automation implementation into an RPA system.

Adaptive automation is another area that could be explored with these different implementations. Adaptive automation takes the basis of automation and adds the ability to change the amount of automation dedicated to each process at any point in time. Task allocation becomes dynamic rather than static, allowing for allocation to change depending on the needs of the operator at a specific point in time. The ultimate goal of adaptive automation is typically to keep the operator from becoming too overworked and/or underworked. With regards to these models, adaptive automation may provide the necessary adjustments to keep the operator engaged but not overworked. It may be able to combine the automation from different stages into a single model to allow for automation to take control of a task at any point. With heightened flexibility, adaptive automation could combine all of the positive factors within each stage to create a system that can best aid the operator in any situation.

Final Conclusions

Many of the results presented above illustrate the diversity of automation implementation. One single type of automation will not be the best solution for every system, which is something designers need to keep in mind when designing automation. The results presented illustrate the effectiveness of automation when implemented in the decision stage with respect to performance. Any designer looking to improve performance may therefore attempt to implement automation at the decision stage for best results. The results also show how the automation can reduce workload drastically when the automation is incorporated in the response stage. Any designer looking to reduce operator workload may therefore attempt to implement automation at the response stage for best results. However, the results suggest the need for further study to determine if these results are specific to the system studied in this research, or if these results are more widely applicable.

Appendix A

Description of Levels and Stages of Automation

L3SA – Computer Offers Alternatives/Information Acquisition

In this combination, the automation will display a set of three different search pattern suggestions using a separate window. The human then decides on one of the search patterns, closes the window with the different search patterns, and the automation completes the search pattern. The human is not required to follow the suggestions of the automation and is only presented with the suggestions. The window appears one time at the beginning of the task. The human cannot decide to view the window again.

Tasks added into the model:

- Display Search Patterns (System task) – this task will take zero seconds to complete. It starts a third path in the model, but only runs once automatically. The human cannot decide to view the window again.
- Decide on Search Pattern (Human task) – this task will take a short amount of time to complete. It is located after the task "Display Search Patterns" and ends the third path. The task uses micromodels Choice Reaction Time (x3), Reading Rate (6 words), Cursor Movement with Mouse (1000 pixels, 200 pixels), and Pushbutton to calculate task time.
- Run Search Pattern (System task) – this task will take the same amount of time as the time it takes to finish the model. It is the last task in the third path, starting

with the task "Display Search Patterns". The human cannot change the search pattern once the path reaches this task.

Tasks changed in the model:

- Find HVT (Human task) – this task will take a reduced amount of time to complete. The search patterns will make the time to find the HVT shorter. Also, the workload will increase overall (not in the task) because the human will have to think about the next step in the search pattern in addition to all of the other workload requirements. Lastly, the task will now start after the task "Decide on Search Pattern". The distribution for the time changes from the original distribution from using the full group of participants. The distribution for the automation is made up of times gathered from three participants that implemented search patterns (subjects 7, 9, and 10).

L5SA – Human Approves Selection/Information Acquisition

In this combination, the automation will decide upon a search pattern and display it through a window. The human will then have the option, within the window, to approve it or deny it. If denied, then the automation will select another search pattern to run. When the human approves the search pattern, the automation will begin to control the camera and complete the search pattern throughout the market while the human attempts to locate the target. At any point, the human can stop the search pattern and take over the automation or request another search pattern.

Tasks added into the model:

- Select Search Pattern (System task) – this task will take zero seconds to complete. It starts a third path in the model, but only runs once automatically. The human can decide to view the window again if desired.
- Approve Search Pattern (Human task) – this task will take a short amount of time to complete. It is located after the task "Select Search Pattern" and may loop back to it based on whether the human approves the selection or not (probability). This task will require a small amount of workload. It will use micromodels Reading Rate (2 words), Simple Reaction Time, On or Off Response, and Cursor Movement with Mouse (1000 pixels, 200 pixels) to calculate task time.
- Run Search Pattern (System task) – this task will take the same amount of time as the time it takes to finish the model. It is the last task in the third path, starting with the task "Select Search Pattern". The human cannot change the search pattern once the path reaches this task.

Tasks changed in the model:

- Find HVT (Human task) – this task will take a reduced amount of time to complete. The search patterns will make the time to find the HVT shorter. Also, the workload will reduce because the human will not have to think about the next step in the search pattern because the search pattern is completed by the automation. The task will now start after the task "Approve Search Pattern". The distribution for the time changes from the original distribution from using the full group of participants. The distribution for the automation is made up of times gathered from three participants that implemented search patterns (subjects 7, 9, and 10).

L7SA – Computer Informs Human of Selection/Information Acquisition

In this combination, the automation will decide upon a search pattern and begin to execute it. A window will appear at the beginning of the task showing which search pattern was chosen, but the human does not have the ability to change the search pattern. At any point, the human may bring up the window to review the search pattern again (making the assumption that they will only need to see it once). Once the task has begun, the automation will control the camera and complete the search pattern throughout the market while the human attempts to locate the target.

Tasks added into the model:

- Run Search Pattern (System task) – this task will take the same amount of time as the time it takes to finish the model. It starts a third path in the model, but only runs once automatically. There is no loop or exit from this task. It is the start of the third path of the model.
- View Search Pattern (Human task) – this task will take a short amount of time to complete. It is located after the task "Run Search Pattern". This task will require a small amount of workload. It will use micromodels Reading Rate (2 words) and Cursor Movement with Mouse (1000 pixels, 200 pixels) to calculate the task time.

Tasks changed in the model

- Find HVT (Human task) – refer to L5SA. The task will now start after the task "View Search Pattern".

L10SA – Full Automation/Information Acquisition

In this combination, the automation will start by running a search pattern. No indicator will appear on the screen to describe the search pattern, so the human does not know which search pattern is being used. Once the task has begun, the automation will control the camera and complete the search pattern throughout the market while the human attempts to locate the target. The human will not have the ability to change which search pattern is being used.

Tasks added into the model:

- Run Search Pattern (System task) – this task will take the same amount of time as the time it takes to finish the model. It starts a third path in the model, but only runs once automatically. There is no loop or exit from this task. It is the only task on the third path of the model.

Tasks changed in the model:

- Find HVT (Human task) – refer to L5SA. In addition, the human will not have to decide on search pattern either, further reducing the workload.

L3TC - Computer Offers Alternatives/Decision and Action Selection

In this combination, the automation will highlight every person in the virtual environment. As the human moves the sensor to different parts of the market, the automation will change the highlight color from green to red when a potential target is identified (person with shovel or weapon). The sensor will need to be zoomed in a certain amount to recognize a potential target enough to change the color from green to red. The human cannot zoom out to view the entire marketplace and allow the automation to pick out the single HVT because the automation uses the same identifiers as the human to identify the HVT. After the HVT has been chosen, all of the highlights go away.

Tasks added in the model:

- Highlight All People (System task) – this task will take zero seconds to complete. It starts a third path in the model, but only runs once automatically. The human cannot ask the automation to re-identify and highlight all of the people in the market again.
- Highlight Potential HVTs (System task) – this task will take zero seconds to complete. It falls on the third path in the model, the next task after the task "Highlight All People". This model will loop back to itself, only active while the human is within the task "Find HVT". The human cannot stop this task from occurring.

Tasks changed in the model:

- Find HVT (Human task) – this task will take a reduced amount of time to complete. By identifying possible HVTs, the automation is removing some of the more obvious distractors and focusing the human attention on certain potential HVTs. The workload does not change because the human is still required to complete the same process of identifying the HVT. The time to complete this task will be based upon the participant times from Scenario 3. Scenario 3 contains a low camera quality and low number of distractors. Highlighting the object carriers in the market will focus the attention of the operator on certain distractors highlighted in red, removing the ones that are only highlighted in green from the decision process of the operator. Scenario 4 with this automation is similar to Scenario 3, so the distribution from "Find HVT" in Scenario 3 will be used.

L5SC - Human Approves Selection/Decision and Action Selection

In this combination, the automation will highlight the single HVT identified with a green color. The human will still need to search the market, but as soon as the HVT is on the screen, the automation will identify the target. The automation will then request confirmation through a pop-up window. The operator will view the identified HVT, and will either accept or reject the identification. If the identification is rejected, then the highlight is removed from the person. If the identification is accepted, then the highlight turns from green to red. The automation will begin the process anew when another HVT appears.

Tasks added into the model:

- Highlight Potential HVT (System task) – this task will take zero seconds to complete. It starts a third path in the model and runs a total of four times. This task is only active while the human is within the task “Find HVT”. The human cannot stop this task from occurring.
- Approve HVT Selection (Human task) – this task will take a short amount of time to complete. It occurs after the task “Highlight Potential HVT” and will loop back to the task “Highlight Potential HVT” when either the human disapproves the selection or the HVT enters the tent and another one appears. Using micromodels Cursor Movement with Mouse (1000 pixels, 200 pixels), Decision Process, Choice Reaction Time (x1), and Mental Rotation/Visualization (0 degrees) to calculate the task time.

Tasks changed in the model:

- Find HVT (Human task) – this task will take a reduced amount of time to complete. By picking out a possible HVT and asking whether the human wants to follow it, the automation is removing all other distractors from the clutter on the screen and focusing the user attention on one single possible target. The workload will not change because the human will still have to decide whether the possible target is the real HVT. For the task time, refer to “Find HVT” in L3TC.

- Lose HVT (Human task) – this task will occur less often. Since each HVT is highlighted and separated from other more obvious distracters, the human will have less of a chance to lose the HVT in the crowd after the HVT has already been identified.

L7SC - Computer Informs Human of Selection/Decision and Action Selection

In this combination, the automation will highlight the single HVT identified with a red color. The human will still need to search the market, but as soon as the HVT is on the screen, the automation will identify the target. The automation will then inform the user of the HVT selection through a pop-up window. Once the HVT has been found, the human will begin to follow the HVT through the market. The process will begin anew when another HVT appears.

Tasks added into the model:

- Highlight HVT (System task) – this task will take zero seconds to complete. This task comes after the task "Find HVT" and before "View Window". The human cannot stop this task from occurring.
- View Window (Human task) – this task will take a small amount of time to complete. It follows the task "Highlight HVT". It contains a small amount of workload to understand what the automation is explaining. It continues with the task "Follow HVT". Using micromodels Cursor Movement with Mouse (1000

pixels, 200 pixels), and Mental Rotation/Visualization (0 degrees) to calculate the task time.

Tasks changed in the model:

- Find HVT (Human task) – this task will take a reduced amount of time to complete. By identifying the HVT when it appears on the screen, the automation is removing all of the possibility of selecting a distractor. The workload decreases because the human now only needs to locate a highlighted target that is selected by the automation. Reference "Find HVT" in L3SC for task time information.
- Lose HVT (Human task) – this task will occur less often. Since each HVT is highlighted and separated from other more obvious distracters, the human will have less of a chance to lose the HVT in the crowd after the HVT has already been identified.

L10SC - Full Automation/Decision and Action Selection

In this combination the automation will highlight the single HVT identified with a red color. The human will still need to search the market, but as soon as the HVT is on the screen, the automation will identify the target. The automation will not inform the user of the target selection. The human will not have the ability to change the HVT once selected.

Tasks added into the model:

- Highlight HVT (System task) – this task will take zero seconds to complete. This task is located after the task “Find HVT” and before the task “Follow HVT”. The human cannot stop this task from occurring.

Tasks changed in the model:

- Find HVT (Human task) – this task will take a reduced amount of time to complete. By identifying the HVT when it appears on the screen, the automation is removing all of the possibility of selecting a distractor. The workload decreases because the human now only needs to locate a highlighted target that is selected by the automation. Reference “Find HVT” in L3SC for task time information.
- Lose HVT (Human task) – this task will occur less often. Since each HVT is highlighted and separated from other more obvious distracters, the human will have less of a chance to lose the HVT in the crowd after the HVT has already been identified.

L3SD - Computer offers alternatives/Action Implementation

In this combination the automation will wait until the F-key is pressed by the human. Once pressed, the automation will request the human to click on the target to follow out of the ones that are on the screen. A pop up window will be used to request identification. Once the target has been decided upon, the automation will take over control of the camera and begin to follow the HVT. The automation will follow the HVT until the HVT enters a tent. During this time, the human will monitor the automation to confirm that the automation is following the target correctly. After that, the human will resume controls and attempt to locate another target within the market. This process will continue until the last HVT enters the tent. If the automation was following a HVT and lost it, the automation will assume that the HVT entered a tent. The operator will be notified that the automation has stopped following the target with a pop up window.

Tasks added into the model:

- Request HVT Selection (System task) – this task will take zero seconds to complete. In the model, it will be located after the task “Find HVT”. It will not require any workload, as it is a system task and not a human task.
- Select HVT (Human task) – this task will take a small amount of time to complete. In the model, it will be located after the task “Request HVT Selection”. It will require a little bit of workload in order to select the HVT on the screen. It will continue on to the “Follow HVT” and “Monitor” tasks after completion. Using micromodels Cursor Movement with Mouse x2 (500 pixels, 100 pixels) (500 pixels, 200 pixels) and Reading Rate (5 words) to calculate task time.
- Monitor (Human task) – this task will take the same amount of time as the task “Follow HVT” to complete. It follows the task “Select HVT”. It will require a small amount of workload to follow the target that the automation is tracking. It does not continue onto anything after completing.
- Notification (Human task) – this task will take a small amount of time to complete. In the model, it will be located after the task “Follow HVT” and before the tasks “Lose HVT” and “HVT in Tent”. It will require a little bit of workload

to read and close the pop up window. Using micromodels Cursor Movement with Mouse (500 pixels, 200 pixels) and Reading Rate (5 words).

Tasks changed in the model:

- Follow HVT (System task) – this task will change from a human task to a system task, removing all of the workload from this task. The amount of time spent in this task will not change.
- Lose HVT (System task) – this task will change from a human task to a system task, removing all of the workload from this task. Because the automation is now following the HVT through the market, the chance that the HVT will be lost depends upon the reliability of the automation in following the HVT.

L5SD - Human Approves Selection/Action Implementation

In this combination the automation will wait until the F-key is pressed by the human. The automation will highlight a specific target and request confirmation from the human that the target highlighted is the one to follow. The request will appear as a pop-up window. The human will accept or deny the request. If denied, then the automation will highlight another target and request confirmation for that target. Once the target has been accepted by the human, the automation will take over control of the camera and begin to follow the HVT. The automation will follow the HVT until the HVT enters a tent. During this time, the human will monitor the automation to confirm that the automation is following the target correctly. After that, the human will resume controls and attempt to locate another target within the market. This process will continue until the last HVT enters the tent. If the automation was following a HVT and lost it, the automation will assume that the HVT entered a tent. The operator will be notified that the automation has stopped following the target with a pop up window.

Tasks added into the model:

- Request HVT Confirmation (System task) – this task will take zero seconds to complete. In the model, it will be located after the task "Find HVT". It will not require any workload, as it is a system task and not a human task.
- Confirm HVT (Human task) – this task will take a small amount of time to complete. In the model, it will be located after the task "Request HVT Confirmation". It will require a little bit of workload in order to confirm the HVT on the screen. It will continue on to the "Follow HVT" and "Monitor" tasks after completion. Using micromodels Cursor Movement with Mouse (500 pixels, 200 pixels), Decision Process, Choice Reaction Time (x1), and Mental Rotation/Visualization (0 degrees) to calculate task time.
- Monitor (Human task) – refer to "Monitor" in L3SD.
- Notification (Human task) – refer to "Notification" in L3SD.
- Reidentify HVT (Human task) – this task will take a small amount of time to complete. In the model, it is located after the task "Lose HVT" and has a single path out of it that continues on to "Request HVT Confirmation". Using micromodels Cursor Movement with Mouse (500 pixels, 1000 pixels), Decision Process, and Pushbutton/Toggle to calculate task time.

Tasks changed in the model:

- Follow HVT (System task) – refer to "Follow HVT" in L3SD.
- Lose HVT (System task) – refer to "Lose HVT" in L3SD.

L7SD - Computer Informs Human of Selection/Action Implementation

In this combination the automation will wait until the F-key is pressed by the human. The automation will highlight a specific target and inform the human that the target highlighted will be followed. The information will appear as a pop-up window. The automation will then take over control of the camera and begin to follow the HVT. The automation will follow the HVT until the HVT enters a tent. During this time, the human will monitor the automation to confirm that the automation is following the target correctly. After that, the human will resume controls and attempt to locate another target within the market. This process will continue until the last HVT enters the tent. If the automation was following a HVT and lost it, the automation will assume that the HVT entered a tent. The operator will be notified that the automation has stopped following the target with a pop up window.

Tasks added into the model:

- Informs of Following (System task) – this task will take zero seconds to complete. In the model, it will be located after the task "Find HVT". It will not require any workload, as it is a system task and not a human task.

- View Window (Human task) – this task will take a small amount of time to complete. It follows the task “Informs of Following”. It contains a small amount of workload to understand what the automation is explaining. It will continue on to the “Follow HVT” and “Monitor” tasks after completion. Using micromodels Cursor Movement with Mouse (500 pixels, 200 pixels), and Mental Rotation/Visualization (0 degrees).
- Monitor (Human task) – refer to “Monitor” in L3SD.
- Notification (Human task) – refer to “Notification” in L3SD.
- Reidentify HVT (Human task) – refer to “Reidentify HVT” in L5SD. After completing, it continues on to “Informs of Following”.

Tasks changed in the model:

- Follow HVT (System task) – refer to “Follow HVT” in L3SD.
- Lose HVT (System task) – refer to “Lose HVT” in L3SD.

L10SD - Full Automation/Action Implementation

In this combination the automation will wait until the F-key is pressed by the human. The automation will then highlight the HVT, take over control of the camera, and begin to follow the HVT. The automation will follow the HVT until the HVT enters a tent. During this time, the human will monitor the automation to confirm that the automation is following the target correctly. After that, the human will resume controls and attempt to locate another target within the market. This process will continue until the last HVT enters the tent. If the automation was following a HVT and lost it, the automation will assume that the HVT entered a tent. The operator will be notified that the automation has stopped following the target with a pop up window.

Tasks added into the model:

- Monitor (Human task) – refer to “Monitor” in L3SD.
- Notification (Human task) – refer to “Notification” in L3SD. The task follows the task “Follow HVT” and continues on to
- Reidentify HVT (Human task) – refer to “Reidentify HVT” in L5SD

Tasks changed in the model:

- Follow HVT (System task) – refer to “Follow HVT” in L3SD.

Lose HVT (System task) – refer to “Lose HVT” in L3SD.

Appendix B

Model Assumptions

Model	Task	Assumption	Assumption Rationale
All Models	HVT Appears	N/A	N/A
All Models	Find HVT	Assumes that the performance values presented from the study are an accurate indication of the amount of time it takes to find a target. The find target time changes as automation is introduced.	This assumption was made because this research assumes that automation may affect the amount of time it takes to find a target.
All Models	Follow HVT	Assumes that the performance values presented from the study are an accurate indication of how long the target was followed, and this value changes as automation is introduced. Also assumes that any time the target is on the screen after the target was found, the operator is following the target.	This assumption was made because this research assumes that automation may affect the amount of time it takes to follow a target.
All Models	Lose HVT	Assumes that the performance values presented from the study are an accurate indication of the amount of time it takes to relocate a target, and this value changes as automation is introduced.	This assumption was made because this research assumes that automation may affect the amount of time it takes to relocate a target.
All Models	Hear Question	Assumes that every question is based on a rectangular distribution from 6.12 sec to 6.50 sec	This assumption was made because the data for the length of the audio recording was unavailable, but an IMPRINT micromodel was used to estimate the amount of time it would take to read the questions out loud.
All Models	Consider Question	Assumes that the performance values presented from the study are an accurate indication of the amount of time it takes to consider the question, and this value does not change as automation is introduced. Also assumes that the entire consider question is spent thinking about the answer to the question.	This assumption was made because this research assumes that automation unrelated to the mathematics question is not going to influence the amount of time to consider the question.
All Models	Respond	Assumes that every answer takes 3 sec to answer. Also assumes that 6% of the questions remian unanswered	This assumption was made because the data for the length of the answering period was unavailable, but an IMPRINT micromodel was used to estimate the amount of time it would take to speak the answer aloud.
All Stage D Models	Monitor	N/A	N/A
All Stage D Models	Notification	Assumes that the time the operator takes to read and close the notification window is based on a rectangular distribution from 1.75 sec to 2.91 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **notify the operator**. More detailed information on the micromodels used can be found in Appendix A.
All Models (Levels 5, 7, 10 in Stage D)	Reidentify HVT	Assumes that the time the operator takes to reidentify the HVT is based on a rectangular distribution from 1.13 sec to 1.88 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **reidentify the HVT**. More detailed information on the micromodels used can be found in Appendix A.

Model	Task	Assumption	Assumption Rationale
Level 3 Stage A	Decide on Search Pattern	Assumes that the time it takes to decide on a search pattern follows a rectangular distribution from 2.96 sec to 4.94 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **decide on a search pattern**. More detailed information on the micromodels used can be found in Appendix A.
Level 3 Stage D	Select HVT	Assumes that the time it takes to select a HVT follows a rectangular distribution from 2.70 sec to 4.50 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **select a HVT**. More detailed information on the micromodels used can be found in Appendix A.
Level 5 Stage A	Approve Search Pattern	Assumes that the time it takes to approve a search pattern follows a rectangular distribution from 1.53 sec to 2.55 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **approve the search pattern**. More detailed information on the micromodels used can be found in Appendix A.
Level 5 Stage C	Approve HVT Selection	Assumes that the time it takes to approve a HVT selection follows a rectangular distribution from 1.87 sec to 3.11 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **approve the HVT selection**. More detailed information on the micromodels used can be found in Appendix A.
Level 5 Stage D	Confirm HVT	Assumes that the time it takes to confirm a HVT follows a rectangular distribution from 1.80 sec to 3.00 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **confirm the HVT selection**. More detailed information on the micromodels used can be found in Appendix A.
Level 7 Stage A	View Search Pattern	Assumes that the time it takes to view a search pattern follows a rectangular distribution from 1.30 sec to 2.16 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **view the search pattern**. More detailed information on the micromodels used can be found in Appendix A.
Level 7 Stage C	View Window	Assumes that the time it takes to view the window follows a rectangular distribution from 1.70 sec to 2.84 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **view the HVT**. More detailed information on the micromodels used can be found in Appendix A.
Level 7 Stage D	View Window	Assumes that the time it takes to view the window follows a rectangular distribution from 1.64 sec to 2.73 sec.	This assumption was made in order to incorporate simulated automation, as this type of automation was not used in the human subject study. IMPRINT micromodels were used to estimate the amount of time to **view the confirmation to follow the HVT**. More detailed information on the micromodels used can be found in Appendix A.
Reliability Models (All Levels in Stages A and C)	Find Failure	Assumes that the operator act of disovering a failure is a random portion of the amount of time that the HVT would be found.	This assumption was made in order to incorporate different reliabilities of simulated automation, as automation and consequently the possibility of failing automation, was not used in the human subject study. A random number between 0-1 was generated and multiplied by a number chosen from the distribution of the HVT find time in the task "Find HVT" to determine the amount of time it took for the human operator to discover the failure.

Model	Task	Processing (Task) Times	Effects	Decision Logic
All Models	HVT Appears	15 sec after the end of the third target; 0 sec every other time	Adds another target to the target counter; resets the time to find the specific target to 0	N/A
All Models	Find HVT	Distribution based on the human subject times	N/A	N/A
All Models	Follow HVT	Distribution based on the human subject times	Calculates the performance score for the specific target	If the task time does not reach the time at which the target enters the tent, the next task is "Lose HVT". Otherwise, the next task is "HVT in Tent".
All Models	Lose HVT	Distribution based on the human subject times	N/A	If the task time does not reach the time at which the target enters the tent, the next task is "Follow HVT". Otherwise, the next task is "HVT in Tent".
All Models	Hear Question	Distribution based on an IMPRINT micromodel	Adds another question to the question counter	N/A
All Models	Consider Question	Distribution based on the human subject times	Calculates the amount of time spent considering the question	N/A
All Models	Respond	Distribution based on an IMPRINT micromodel	Calculates the communication score for the specific question	If the fourth question has been asked, then there is no further task. Otherwise, the next task is "Question Delay".
All Stage D Models	Monitor	Amount of time that remains to follow the specific HVT	N/A	N/A
All Stage D Models	Notification	Distribution based on IMPRINT micromodels	N/A	If the task time does not reach the time at which the target enters the tent, the next task is "Lose HVT". Otherwise, the next task is "HVT in Tent".
All Models (Levels 5, 7, 10 in Stage D)	Reidentify HVT	Distribution based on IMPRINT micromodels	N/A	N/A
Level 3 Stage A	Decide on Search Pattern	Distribution based on IMPRINT micromodels	N/A	N/A
Level 3 Stage D	Select HVT	Distribution based on IMPRINT micromodels	Calculates how long the operator will have to find the target after selecting a HVT	N/A
Level 5 Stage A	Approve Search Pattern	Distribution based on IMPRINT micromodels	Updates model to include that a search pattern has been approved	If the search pattern has been approved, the next task is "Run Search Pattern". Otherwise, the next task is "Select Search Pattern".
Level 5 Stage C	Approve HVT Selection	Distribution based on IMPRINT micromodels	N/A	N/A
Level 5 Stage D	Confirm HVT	Distribution based on IMPRINT micromodels	Calculates how long the operator will have to find the target after confirming a HVT	N/A
Level 7 Stage A	View Search Pattern	Distribution based on IMPRINT micromodels	N/A	N/A
Level 7 Stage C	View Window	Distribution based on IMPRINT micromodels	N/A	N/A
Level 7 Stage D	View Window	Distribution based on IMPRINT micromodels	Calculates how long the operator will have to find the target after confirming a HVT	N/A
Reliability Models (All Levels in Stages A and C)	Find Failure	Distribution based on distribution used in the task "Find HVT"	N/A	N/A
Notes:	All system tasks are not included because they are assumed to be unaffected by any change in the automation or relibility.			

Bibliography

Alion Science and Technology. (2009). *IMPRINT Pro User Guide.* Boulder, Colorado: Alion Science and Technology.

Bainbridge, L. (1983). Ironies of Automation. *Automatica*, 775-779.

Breslow, L. A., Gartenberg, D., McCurry, M., & Trafton, G. (2014). Dynamic Operator Overload: A Model for Predicting Workload During Supervisory Control. *IEEE Transactions on Human-Machine Systems*, 30-40.

Broadbent, D. E. (1958). *Perception and Communication.* London, U.K.: Pergamon.

Callam, A. (2014). Drone wars: Armed unmanned aerial vehicles. *International Affairs Review*.

Clare, A. S., Hart, C. S., & Cummings, M. L. (2010). Assessing operator workload and performance in expeditionary multiple unmanned vehicle control. *AIAA Aerospace Sciences Meeting.* Orlando.

Colombi, J. M., Miller, M. E., Schneider, M., McGrogan, J., Long, D. S., & Plaga, J. (2012). Predictive mental workload modeling for semiautonomous system design: Implications for systems of systems. *Systems Engineering*, 448-460.

Cring, E. A., & Lenfestey, A. G. (2009). *Architecting human operator trust in automation to improve system effectiveness in multiple unmanned aerial vehicles (UAV) control.* AFIT.

Cummings, M. L., Bruni, S., Mercier, S., & Mitchell, P. J. (2007). *Automation architecture for single operator, multiple UAV command and control.* Cambridge: Massachusets Institute of Technology.

De Visser, E. J., LeGoullon, M., Freedy, A., Freedy, E., Weltman, G., & Parasuraman, R. (2008). Designing an adaptive automation system for human supervision of unmanned vehicles: A bridge from theory to practice. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 221-225.

Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission Control of Multiple Unmanned Aerial Vehicles: A Workload Analysis. *Human Factors*, 479-487.

Donmez, B., Nehme, C., & Cummings, M. L. (2010). Modeling workload impact in multiple unmanned vehicel supervisory control. *Systems, Man and Cybernetics*, 1180-1190.

Endsley, M. R. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 462-492.

Fitts, P. M. (1951). *Human engineering for an effective air-navigation and traffic-control system.*

Furukawa, H., Inagaki, T., & Niwa, Y. (2000). Operator's situation awareness under different levels of automation; Evaluations through probabilistic human cognitive simulations. *Systems, Man, and Cybernetics, 2000 IEEE International Conference* (pp. 1319-1324). IEEE.

Gertler, J. (2012). *U.S. Unmanned Aerial Systems.* Washington D.C.: Congressional Research Service.

Hancock, P. A. (2007). On the process of automation transition in multitask human-machine systems. *IEEE Transactions on Systems, Man and Cybernetics,* 586-598.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, 139-183.

Hoc, J.-M. (2001). Towards a cognitive approach to human-machine cooperation in dynamic situations. *International Journal of Human-Computer Studies*, 509-540.

Howard, A. M. (2007). A Systematic Approach to Predict Performance of Human-Automation Systems. *IEEE Transactions on Systems, Man & Cybernetics,* 594.

Inagaki, T. (2003). Adaptive Automation: Sharing and Trading of Control. In E. Hollnagel, *Handbook of Cognitive Task Design* (pp. 147-169). Mahwah: Lawrence Erlbaum Associates.

Johansson, B., Fasth, A., Stahre, J., Heilala, J., Leong, S., Lee, Y. T., & Riddick, F. (2009). Enabling Flexible Manufacturing Systems by Using Level of Automation as Design Parameter. *Proceedings of the 2009 Winter Simulation Conference* (pp. 2176-2184). IEEE.

John, M. S., & Manes, D. I. (2002). Making unreliable automation useful. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 332-336). SAGE Publications.

Jones, S. (2014, October 12). *Drone strikes on Isis loom large in allies' strategy*.

Kaber, D. B., Stoll, N., & Thurow, K. (2007). Human-automation interaction strategies for life science applications: Implications and future research. *IEEE International Conference on IEEE.*

Kaber, D. B., Stoll, N., & Thurow, K. (2007). Human-automation interaction strategies for life sciences applications. Implications and future research. *IEEE International Conference* (pp. 615-620). IEEE.

Keller, J. (2002). Human performance modeling for discrete-event simulation: workload. *Proceedings of the Winter* (pp. 157-162). IEEE.

McCracken, J. H., & Aldrich, T. B. (1984). *Analyses of selected LHX mission functions: Implications for operator workload and system automation goals.* Fort Rucker: Anacapa Sciences.

Meyer, J., Feinshreiber, L., & Parmet, Y. (2003). Levels of automation in a simulated failure detection task. *Systems, Man and Cybernetics* (pp. 2101-2106). IEEE.

Misener, P. (2014, July 9). Amazon Petetion for Exemption. Seattle, Washington: Amazon.com.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors,* 476.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced "Complacency". *The International Journal of Aviation Psychology*, 1-23.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics*, 12.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 76-87.

Ruff, H. A., Calhoun, G. L., Draper, M. H., Fontejon, J. V., & Guilfoos, B. J. (2004). *Exploring automation issues in supervisory control of multiple UAVs.* Dayton: Sytronics.

Rusnock, C. F. (2013). Simulation-based Cognitive Workload Modeling and Evaluation of Adaptive Automation Invoking and Revoking Strategies. Orlando: University of Central Florida.

Rusnock, C. F., & Geiger, C. D. (2014). Simulation-based Assessment of Performance-Workload Tradeoffs for System Design Evaluation. *Industrial and Systems Engineering Research Conference.* Montreal: ISERC.

Sheridan, T. B. (2011). Adaptive Automation, Level of Automation, Allocation Authority, Supervisory Control, and Adaptive Control: Distinctions and Modes of Adaptation. *IEEE Transactions on Systems, Man, and Cybernetics* (pp. 662-667). IEEE.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators.* Cambridge: Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Swanson, L., Jones, E., Riordan, B., Bruni, S., Schurr, N., Sullivan, S., & Lansey, J. (2012). Exploring human error in an RPA target detection task. *Human Factors and Ergonomics Society Annual Meeting* (pp. 328-332). Sage Publications.

Taylor, R. M. (2006). *Human automation integration for supervisory control of UAVs.* Farnborough: Defence Science and Technology Lab.

Tvaryanas, A. P., Thompson, W. T., & Constable, S. H. (2006). Human factors in remotely piloted aircraft operations: HFACS analysis of 221 mishaps over 10 years. *Aviation, space, and environmental medicine*, 724-732.

Van Cleave, D. W. (2003). Trends and Technologies for Unmanned Aerial Vehicles. In T. Samad, & G. Balas, *Software-Enabled Control: Information Technology for Dynamical Systems* (pp. 9-26). Hoboken: Institute of Electrical and Electronics Engineer.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman, & D. R. Davies (Eds.), *Varieties of attention.* San Diego.

Wickens, C. D., & Dixon, S. R. (2006). Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload. *Human Factors*, 474-486.

Woods, D. D., Johannesen, L. J., Cook, R. I., & Sarter, N. B. (1994). *Behind human error: Cognitive systems, computers and hindsight.* Dayton, OH: University of Dayton Research Institute.

Wright, M. C., & Kaber, D. B. (2005). Effects of automation of information-processing functions on teamwork. *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 50-66.

Wu, C., & Yili, L. (2009). Development and evaluation of an ergonomic software package for predictiing multiple-task human performance and mental workload in human-machine interface design and evaluation. *Computers & Industrial Engineering*, 323-333.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and physchology*, 459-482.

REPORT DOCUMENTATION PAGE			*Form Approved* *OMB No. 074-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
26-03-2015	Master's Thesis	October 2013 – March 2015

TITLE AND SUBTITLE	
The Effect of Stages and Levels of Automation and Reliability on Workload and Performance for Remotely Piloted Aircraft Operations	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	**5d. PROJECT NUMBER**
Katrein, Stephen P., 2d Lt, USAF	**5e. TASK NUMBER**
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)	**8. PERFORMING ORGANIZATION REPORT NUMBER**
Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way, Building 640 WPAFB OH 45433-8865	AFIT-ENV-MS-15-M-201
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)** **711th HPW/RHCP**
Scott Galster 711th Human Performance Wing 2947 Fifth Street, WPAFB OH 45433 Scott.galster@us.af.mil	**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT
This thesis investigates incorporating different stages and levels of automation with varying degrees of reliability into a remotely piloted aircraft (RPA) surveillance task in order to determine how automation implementation and reliability affect operator workload and system performance. The study uses IMPRINT discrete event simulation to evaluate three levels of reliability in twelve different baseline automation implementations within a remotely piloted vehicle task. Three stages and four levels are modeled, for a total of twelve combinations, along with a baseline task with no automation. The stages modeled are the information acquisition stage, the decision and action selection stage, and the action implementation stage, coupled with the automation recommendation level, the operator consent level, the operator veto level, and the fully automatic level. The reliability is assessed at 100%, with reduced reliabilities of 80%, 70%, and 60%. This study finds that stages of automation have greater impact on performance and the workload values than levels of automation. Automation with reduced reliability is found to have significantly reduced performance for all stages except the response stage models. However, reductions in reliability are found to have little impact on operator workload.

15. SUBJECT TERMS
Stages and Levels of Automation, DES, IMPRINT, Reliability, Workload, Performance

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Christina Rusnock, AFIT/ENV
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include area code) (937) 255-3636, ext 4611 christina.rusnock@afit.edu
U	U	U	UU	148	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18